Signature moments to characterize laws of stochastic processes

SIGNATURE MOMENTS TO CHARACTERIZE LAWS OF STOCHASTIC PROCESSES

ILYA CHEVYREV AND HARALD OBERHAUSER

Abstract. The normalized sequence of moments characterizes the law of any finite-dimensional random variable. We prove an analogous result for path-valued random variables, that is stochastic processes, by using the normalized sequence of signature moments. We use this to define a metric for laws of stochastic processes. This metric can be efficiently estimated from finite samples, even if the stochastic processes themselves evolve in high-dimensional state spaces. As an application, we provide a non-parametric two-sample hypothesis test for laws of stochastic processes.

1. Introduction

Moments of vector-valued data. Let X be an $\mathbb{R}^d$-valued random variable and denote with $\mu_X = \mathbb{P} \circ X^{-1}$ the law of X. There are several ways to characterise μ_X among all finite measures on $\mathbb{R}^d$:

(i) the sequence of moments

$$\left(\mathbb{E}\left[X^{\otimes m}\right]\right)_{m\geq 0} \in \mathbf{T}(\mathbb{R}^d) := \prod_{m\geq 0} \left(\mathbb{R}^d\right)^{\otimes m},$$

(ii) the Fourier transform (complex moments)

$$\mathbb{R}^d \ni u \mapsto \mathbb{E}[e^{i\langle u, X\rangle}] \in \mathbb{C}\,,$$

(iii) the sequence of normalised moments

$$\left(\mathbb{E}\left[\lambda(X)^m X^{\otimes m}\right]\right)_{m\geq 0} \in \mathbf{T}(\mathbb{R}^d)\,,$$

where $\lambda : \mathbb{R}^d \to (0, \infty)$ is a suitable normalising function.

Approach (i) requires that the sequence $(\mathbb{E}[X^{\otimes m}])_{m\geq 0}$ does not grow too fast. Approaches (ii) and (iii) characterize μ_X without any assumptions on the random variable X. While (iii) is a less well-known generalization of (i), it follows from a simple argument (we give full details in Section 4).

The above elementary facts about laws of random variables in finite-dimensional spaces are at the core of many results in statistics and machine learning. In this article we study the infinite-dimensional case of laws of path-valued random variables — that is, when $X = (X_t)$ is a stochastic process — by replacing monomials of a vector by iterated integrals of paths; this collection of integrals is knowns as the path signature in stochastic analysis. We then use this to define a metric for measures on pathspace, study the topology it generates, and give applications to hypothesis testing.

Path-valued data. For many real-world data sets one sample/observation $X(\omega) = x = (x_t)_{t\in[0,1]} \in C([0, 1], \mathbb{R}^d)$ is a path (e.g. patient data, speech signals, finance). A few important points specific to path-valued data are

time parameterization (in)invariance: for many real-world signals, it is beneficial to ignore parameterization of time; e.g. in speech, different speakers pronounce the same words at different speeds.

discrete observations: usually only discrete time observations $(x(t_i))_{i=1,\dots,n}$ are provided (due to finite storage, sampling cost, etc). However, treating this as a learning problem of vector-valued data is problematic since the number n and position of time points might change from sample to sample. Additionally, n can get very large for high-frequency data.

unbounded variation paths: unbounded variation paths arise by Donsker-type theorems in the high-frequency limit and functions of such paths have to be treated with care. For example, important functions such as quadratic variation, or the solution of nonlinear filtering or a stochastic differential equation, do not depend continuously on the underlying path.

Our approach relies on a feature map from stochastic analysis called the "signature" of a path which addresses all three issues: it is parameterization invariant (with the option to capture parameterization variance); it is robust to irregular sampling a path at irregular times; it appears naturally when describing the behaviour of functions of non-smooth paths.

Mathematical Institute, University of Oxford, Andrew Wiles Building, Woodstock Road, Oxford OX2 6GG, United Kingdom
E-mail addresses: chevyrev@maths.ox.ac.uk, oberhauser@maths.ox.ac.uk.

IC is supported by a Junior Research Fellowship of St John's College, Oxford.
HO is supported by the Oxford-Man Institute of Quantitative Finance.

Moments of path-valued data. The starting point of our investigation is a classical theme from stochastic analysis: if $x = (x_t)_{t\in[0,1]} \in C([0,1],\mathbb{R}^d)$ is a path, then

$$\int dx^{\otimes m} := \int_{0<t_1<\cdots<t_m<1} dx(t_1)\otimes\cdots\otimes dx(t_m) \tag{1}$$

behaves in a precise algebraic sense analogous to a monomial of degree m of an element in $\mathbb{R}^d$. The so-called signature map

$$x \mapsto \left(\int dx^{\otimes m}\right)_{m\geq 0}$$

is injective up to tree-like equivalence.[1] The iterated integrals above are defined through either classical Riemann-Stieltjes integrals, or as stochastic or rough path integrals if x is not smooth. Since these iterated integrals behave like monomials, one can hope to extend classical results characterizing laws from the finite dimensional case to pathspace.

Example 1.1. *If $X \sim N(0, I_d)$, the standard multivariate normal in $\mathbb{R}^d$, then*

$$\left(\mathbb{E}\left[X^{\otimes m}\right]\right)_{m\geq 0} = \exp\left(\frac{1}{2}I_d\right)$$

and the classical Carleman's condition implies that this sequence uniquely determines the law of X. Brownian motion in $\mathbb{R}^d$ is arguably the natural, infinite-dimensional analogue to a multivariate normal; indeed a direct calculation gives the analogous formula

$$\left(\mathbb{E}\Big[\int dX^{\otimes m}\Big]\right)_{m\geq 0} = \exp\Big(\frac{1}{2}\sum_{i=1}^{d} e_i\otimes e_i\Big), \tag{2}$$

where X denotes a standard Brownian motion and the stochastic integrals are taken in the Stratonovich sense.

While in the finite dimensional case it is well-understood how moments describe the law of a random variable — see Points (i), (ii), (iii) above — this question is more subtle in the infinite-dimensional case of path-valued random variables,variables, one of the difficulties being that the pathspace $C([0,1],\mathbb{R}^d)$ is not locally compact.

(I) Inspired by (i), Fawcett [19] showed that

$$\left(\mathbb{E}\Big[\int dX^{\otimes m}\Big]\right)_{m\geq 0} \in \mathbf{T}(\mathbb{R}^d)$$

characterizes the law of the stochastic process $X = (X_t)$ up to tree-like equivalence, provided this law has compact support. However, the assumption of compact support is usually much too strong; for example, it certainly does not apply to Brownian motion in Example 1.1.

(II) Inspired by (ii), a (non-commutative) extension of the Fourier transform to pathspace was developed in [15]. This Fourier transform characterises laws of any stochastic process $X = (X_t)$ up to tree-like equivalence.[2] However, it is quite abstract and has so far been elusive to concrete computations.

(III) Inspired by (iii), we show that for suitable normalizations λ, the sequence

$$\left(\mathbb{E}\Big[\lambda(X)^m\int dX^{\otimes m}\Big]\right)_{m\geq 0}$$

characterizes the law of $X = (X_t)$ up to tree-like equivalence. The only assumption on X is that the iterated integrals are well-defined. Unlike (II), this leads to efficient algorithms that can be combined with well-developed tools from machine learning such as maximum mean distances and kernelization.

The proof of (III) relies on combining the (dual) notion of a universal and characteristic feature map with ideas from stochastic analysis. A simple variation (adding time as a component), allows to characterize the law of X, i.e. also distinguish tree-like equivalent processes (such as processes that differ only by a time change). Below we sketch the main ideas and applications.

[1] Tree-like equivalence is a very useful equivalence relation on pathspace; e.g. it identifies paths up to a time-change. From an analytic point of view, tree-like equivalence is the analogous notion of Lebesgue almost sure equivalence of sets in $\mathbb{R}^d$ on pathspace.

[2] In general, it gives a much more complete picture than (I); for example, it implies that (2) characterizes the law of Brownian motion. This applies to many other processes under suitable growth assumptions.

Universal and characteristic features. Much of statistical learning theory relies on finding a feature map that embeds the data — in our case, samples from a stochastic process — into a linear space. Two requirements for a "good" feature map are

universality: non-linear functions of the data are approximated by linear functionals in feature space,
characteristicness: the expected value of the feature map characterizes the law of the random variable.

A useful observation is that these two properties are in duality and therefore often equivalent [46].[3]

Our feature map will be of the form

$$\Phi : x \mapsto \left(\lambda(x)^m \int dx^{\otimes m}\right)_{m\geq 0}$$

taking an unparameterized path $x = (x_t)$ to an element in the linear space $\mathbf{T}(\mathbb{R}^d)$. To show "characteristicness" of Φ, it is, by the above duality, sufficient to show "universality". For the latter, we use a convenient generalization of the Stone–Weierstrass theorem together with the recently established injectivity of the map $x \mapsto \left(\int dx^{\otimes m}\right)_{m\geq 0}$ in [4].

Maximum mean distances. A natural distance between probability measures is the "maximum mean distance"

$$d(\mu, \nu) = \sup_f \left|\mathbb{E}_{X\sim\mu}[f(X)] - \mathbb{E}_{Y\sim\nu}[f(Y)]\right|. \tag{3}$$

where the sup is taken over a sufficiently large set of real-valued functions. To approximate (3) from finite samples of the laws μ resp. ν of stochastic processes X resp. Y, naive approaches are troublesome because of the supremum over a large space of functions. To address this, we follow [2, 26, 47, 33] and kernelize the signature feature map Φ. That is, we define the signature kernel

$$\mathrm{k}(x, y) := \langle \Phi(x), \Phi(y) \rangle$$

and take the sup in (3) over functions in the unit ball of the reproducing kernel Hilbert space (henceforth RKHS) of k. This gives the identity

$$d_{\mathrm{k}}(\mu, \nu) = \mathbb{E}[\mathrm{k}(X, X')] - 2\mathbb{E}[\mathrm{k}(X, Y)] + \mathbb{E}[\mathrm{k}(Y, Y')] , \tag{4}$$

where X' and Y' denote independent copies of $X \sim \mu$ and $Y \sim \nu$. Efficient recursive algorithms for the signature k have been developed [33] and this allows to evaluate $d_k(\mu, \nu)$ from finite samples, even if the paths X,Y evolve in high-dimensional spaces (large d) or general topological spaces.

Topologies induced by maximum mean distances are typically hard to relate to standard topologies, see [42]. However, we show that d_{k} induces a topology strictly weaker than classical weak convergence (also called narrow convergence/convergence in law). This also answers a question about learning on non-compact spaces that was recently raised in [46].[4]

Application: hypothesis testing for laws of stochastic processes. We apply our theoretical results in the context of two-sample hypothesis testing [26, 27] for stochastic processes. A two-sample test for stochastic processes $X = (X_t)_{t\in[0,1]}$, $Y = (Y_t)_{t\in[0,1]}$, tests the null-hypothesis

$$H_0 : P_X = P_Y \text{ against the alternative } H_1 : P_X \neq P_Y,$$

where $P_X := \mathbb{P} \circ X^{-1}$ and $P_Y := \mathbb{P} \circ Y^{-1}$ are the laws of X and Y. We have implemented our method and present two elementary but natural examples from stochastic processes for which approaches that identify path samples as large vectors become very inefficient.

Summary, outline, and notation. Our main theoretical results are

(a) a characteristic and universal feature map, $x \mapsto \Phi(x)$, for paths,
(b) a characteristic and universal kernel $\mathrm{k}(x, y) = \langle \Phi(x), \Phi(y) \rangle$ for paths,
(c) a metric for laws of unparameterized stochastic processes,

together with efficient algorithms that allow to use these results, even if the paths evolve in high-dimensional or non-linear spaces. In fact, we provide two feature maps: one that is invariant to time-parameterization and one that is not. As an application, we give a new, non-parametric two-sample hypothesis test for laws of stochastic processes.

We collect some commonly used notation in the following table.

[3]The subtlety is that probability measures form a convex space but not a linear space, so one has to work with general distributions rather than probability measures for characteristicness.

[4]Our construction provides a negative answer to the question raised in [46] asking whether every bounded, continuous kernel over a Polish (non-locally compact) space which is characteristic to probability measures metrizes weak convergence. For sake of clarity, we also provide a direct counterexample in Proposition 7.6 and Appendix F, which is independent of our construction on pathspace.

Symbol	Meaning	Page
	Spaces	
$\mathcal{X}$	topological space	
E	topological vector space (TVS) over $\mathbb{R}$	
E^*	algebraic dual of E, i.e., space of linear functionals $E \to \mathbb{R}$	
E'	topological dual of E	
H	Hilbert space over $\mathbb{R}$	
V	Banach space over $\mathbb{R}$	
	Paths and sequences	
C^1	subset of $C([0,1],V)$ of bounded variation paths	8
$\mathcal{P}^1$	tree-like equivalence classes of bounded variation paths in V	23
$x \sim_t y$	tree-like equivalence relation between $x, y \in C([0,1],V)$	8
π	partition of $[0,1]$, i.e. a collection of points $0 \le t_1 < \cdots < t_l \le 1$	
x^π	the sequence $x^\pi := (x(t_i))_{i=1,\dots,l}$ given by sampling $x = (x(t))_{t\in[0,1]}$ along π	
	Signatures and normalization	
$\int dx^{\otimes m}$	shorthand for $\int_{0\le t_1\le\cdots\le t_m\le 1} dx(t_1)\otimes\cdots\otimes dx(t_m)$	8
S	signature map, $\mathrm{S}(x) = \left(\int dx^{\otimes m}\right)_{m\ge 0}$	8
$\mathbf{T}(V)$	tensor algebra over V	6
$\mathbf{T}_1(V)$	subset of $\mathbf{T}(V)$ with zero-th component equal 1	6
Λ	a tensor normalization	7

The paper is organised as follows. Section 2 introduces the equivalence between universality and characteristicness of a feature map $\Phi : \mathcal{X} \to E$, as well as the strict topology that becomes useful when $\mathcal{X}$ is non-compact. Section 3 focuses on the tensor algebra, $E := \mathbf{T}(V)$, as the natural feature space for monomials and introduces a tensor normalization $\Lambda : \mathbf{T}_1(V) \to \mathbf{T}_1(V)$. Section 4 applies the results from Sections 2 and Section 3 to give an elementary proof of the characterization (iii) of laws of finite-dimensional random variables. Section 5 and 6 then uses the normalized signature map $\Phi = \Lambda \circ \mathrm{S}$ to extend this argument from finite-dimensional random variables to path-valued random variables. Section 7 introduces a bounded, universal and characteristic kernel $\mathrm{k}(x,y) = \langle \Phi(x), \Phi(y)\rangle$ for paths x, y and discusses the associated MMD and the topology it induces on the space of laws of stochastic processes. Section 8 contains as application a two-sample test between laws of stochastic processes and numerical experiments on two elementary examples: testing whether a signal or pure noise is observed, and testing whether a simple random walk is observed.

Remark 1.2. *Throughout the main text we focus on readability over generality and postpone many proofs and generalizations to the appendix. For example, our results about characteristicness of the normalized signature Φ extend from the domain of bounded variation paths C^1 (resp. tree-like equivalent bounded variation paths $\mathcal{P}^1$) to geometric as well as branched p-rough paths for any $p \ge 1$ that evolve in general Banach spaces V. Leaving applications in statistics aside, these are non-trivial results in stochastic analysis and we give the statements in full generality with complete proofs in the appendix.*

2. Learning in non-locally compact spaces

Two classical problems of learning in a topological space $\mathcal{X}$ are to

(1) make inference about a function $f \in \mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$,
(2) make inference about a probability measure μ on $\mathcal{X}$.

The standard approach in statistical learning [18] is to map $\mathcal{X}$ into a (typically infinite or high dimensional) linear space E and address the learning problem there by using linear methods.

Definition 2.1. *Let $\mathcal{X}$ be a topological space and $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ a TVS. We call $\mathcal{X}$ the input space and $\mathcal{F}$ the hypothesis space. For a TVS E, we call any map*

$$\Phi : \mathcal{X} \to E$$

a feature map for which E is the feature space.

2.1. **Universal and characteristic features.** To address Point (1), we require that E' is large enough to approximate the elements of the function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, i.e. for every $f \in \mathcal{F}$, there exists $\ell \in E'$ such that

$$f(\cdot) \approx \langle \ell, \Phi(\cdot) \rangle$$

as real-valued functions on $\mathcal{X}$. To address Point (2), we require that the feature map is non-linear enough to distinguish measures μ on $\mathcal{X}$, i.e. the map

$$\mu \mapsto \mu(\Phi) := \int_{\mathcal{X}} \Phi(x)\mu(dx) \in E$$

is injective. Since E can be infinite dimensional, it is more convenient to write this map as

$$\mu \mapsto \Big(\ell \mapsto \int_{\mathcal{X}} \langle \ell, \Phi(x) \rangle \mu(dx)\Big) \in (E')^* \ .$$

More generally, we can replace the integral by any distribution $D \in \mathcal{F}'$, i.e. require that

$$D \mapsto (\ell \mapsto D(x \mapsto \langle \ell, \Phi(x) \rangle))$$

is injective. The definition of universality and characteristicness of Φ makes this precise.

Definition 2.2. *Fix an input space $\mathcal{X}$ and hypothesis space $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$. Consider a feature map*

$$\Phi : \mathcal{X} \to E \ .$$

Suppose that $\langle \ell, \Phi(\cdot) \rangle \in \mathcal{F}$ for all $\ell \in E'$. We say that Φ is

(a) universal to $\mathcal{F}$ if the map

$$\iota : E' \to \mathbb{R}^{\mathcal{X}}, \quad \ell \mapsto \langle \ell, \Phi(\cdot) \rangle \tag{5}$$

has a dense image in $\mathcal{F}$.

(b) characteristic to a subset $\mathcal{P} \subset \mathcal{F}'$ if the map

$$\kappa : \mathcal{P} \to (E')^*, \quad D \mapsto [\ell \mapsto D(\langle \ell, \Phi(\cdot) \rangle)]$$

is injective.

Universality directly addresses Point (1); characteristicness addresses Point (2) in a much more general sense since the dual $\mathcal{F}'$ is typically much larger than the set of probability measures on $\mathcal{X}$. This generalization allows to work on the linear space $\mathcal{F}'$ instead of the convex subset of probability measures. A consequence is the following duality between universality and characteristicness.

Theorem 2.3. *Suppose that $\mathcal{F}$ is a locally convex TVS. A feature map Φ is universal to $\mathcal{F}$ iff Φ is characteristic to $\mathcal{F}'$.*

Proof. First assume Φ is universal to $\mathcal{F}$. Since the image of ι is dense in $\mathcal{F}$, every element $D \in \mathcal{F}'$ is determined by its values on the set $(\langle \ell, \Phi(\cdot) \rangle)_{\ell \in E'} \subset \mathcal{F}$, hence Φ is characteristic to $\mathcal{F}'$. Now assume Φ is not universal. By the Hahn–Banach Theorem, there exists $\ell \in \mathcal{F}'$ such that $\ell \neq 0$ but $\ell|_{\iota(E)} = 0$. In particular, $\kappa(\ell) = 0$, hence κ is not injective. □

Theorem 2.3 is a reformulation of a corresponding result for kernels from [46]. Although it follows from an elementary application of Hahn–Banach, we will see that it is useful since universality is often much easier to show than characteristicness.

Example 2.4. *Let $\mathcal{X} \subset \mathbb{R}^d$ compact and $E := \prod_{m \geq 0} \left(\mathbb{R}^d\right)^{\otimes m}$ equipped with the product topology. Consider the feature map*

$$\Phi : \mathcal{X} \to E \ , \quad \Phi(x) := \Big(\frac{x^{\otimes m}}{m!}\Big)_{m \geq 0}.$$

The image $\iota(E')$ is the space of polynomials on $\mathcal{X}$. By Stone–Weierstrass, it follows that Φ is universal to $\mathcal{F} := C(\mathcal{X}, \mathbb{R})$ equipped with the uniform topology. By Theorem 2.3, Φ is characteristic to $\mathcal{F}'$, which is the set of regular Borel measures on $\mathcal{X}$, i.e., the map

$$\mu \mapsto \Big(\int_{\mathcal{X}} \frac{x^{\otimes m}}{m!} \mu(dx) \Big)_{m \geq 0}$$

is injective, which agrees with the well-known fact that a regular Borel measure on compacts is completely determined by its moments.

2.2. **The strict topology.** To extend Example 2.4 from compact to non-compact $\mathcal{X}$, we need to

(a) find a map $\Phi : \mathcal{X} \to E$ that behaves like monomials,
(b) find a TVS $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ such that $\mathcal{F}'$ includes the probability measures on $\mathcal{X}$,
(c) have a Stone–Weierstrass-type result for (a subset of) $\{x \mapsto \langle \ell, \Phi(x)\rangle, \ell \in E'\}$.

Point (a) must take into account the concrete choice of $\mathcal{X}$; for the case when $\mathcal{X}$ is a pathspace, we discuss in detail in Section 5 how the (normalized) signature has these properties. Here, we address points (b) and (c) in full generality. Note that for non-locally compact spaces $\mathcal{X}$ it is often easy to find $\mathcal{F}$ such that exactly one of (b) and (c) holds, but not both; typically the dual is either too small, or Stone–Weierstrass-type results are not known or involve conditions that are hard to verify. Below we recall a result of Giles [25] which determines a convenient choice of $\mathcal{F}$ with the above properties.

Definition 2.5. *Let $\mathcal{X}$ be a topological space. We say that a function $\psi : \mathcal{X} \to \mathbb{R}$ vanishes at infinity if for all $\epsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that $\sup_{x\in\mathcal{X}\setminus K} |\psi(x)| < \epsilon$. Denote with $B_0(\mathcal{X}, \mathbb{R})$ the set of functions that vanish at infinity. The strict topology[5] on $C_b(\mathcal{X}, \mathbb{R})$ is the topology generated by the seminorms*

$$p_\psi(f) = \sup_{x\in\mathcal{X}} |f(x)\psi(x)| , \quad \psi \in B_0(\mathcal{X}, \mathbb{R}) .$$

Theorem 2.6 (Giles [25])**.** *Let $\mathcal{X}$ be a metrizable topological space.*

(1) The strict topology on $C_b(\mathcal{X}, \mathbb{R})$ is weaker than the uniform topology and stronger than the topology of uniform convergence on compact sets.
(2) If $\mathcal{F}_0$ is a subalgebra of $C_b(\mathcal{X}, \mathbb{R})$ such that
  (a) $\forall x, y \in \mathcal{X}$ there exists $f \in \mathcal{F}_0$ such that $f(x) \neq f(y)$,
  (b) $\forall x \in \mathcal{X}$ there exists a $f \in \mathcal{F}_0$ such that $f(x) \neq 0$,
  then $\mathcal{F}_0$ is dense in $C_b(\mathcal{X}, \mathbb{R})$ under the strict topology.
(3) The topological dual of $C_b(\mathcal{X}, \mathbb{R})$ equipped with the strict topology is the space of finite regular Borel measures on $\mathcal{X}$.

Proof. Point (1) follows from the definition. Point (2) is [25, Thm. 3.1]; in fact, the result holds more generally for any topological space and one only needs that point-separating and non-vanishing functions are in the closure of $\mathcal{F}_0$. Point (3) is [25, Thm. 4.6] (which applies more generally to k-spaces). □

Remark 2.7.

- *Another option is to equip $\mathcal{F} = C_b(\mathcal{X}, \mathbb{R})$ with the topology of uniform convergence, so that the topological dual $\mathcal{F}' = \mathrm{rba}(\mathcal{X})$ is the space of regular bounded finitely additive measures on $\mathcal{X}$. A variant of Stone–Weierstrass holds but it comes with assumptions that are often hard to verify since they involve ''topological zero sets", see* [48, Prob. 44A].
- *Yet another option is to use a (e.g. Stone–Čech) compactification $\beta\mathcal{X}$ of $\mathcal{X}$. Then $\mathcal{F}'$ can be identified with the space of regular Borel measures on $\beta\mathcal{X}$. Unfortunately, this leads in general to very abstract objects.*

3. Tensor algebras and normalization

We discuss in this section the feature space in which our feature map will take values. If V is a vector space, then $\prod_{m\geq 0} V^{\otimes m}$ is a linear space by extending tensor addition by linearity

$$\mathbf{s} + \mathbf{t} := \left(\mathbf{s}^0 + \mathbf{t}^0, \mathbf{s}^1 + \mathbf{t}^1, \ldots\right) \text{ for } \mathbf{s} = (\mathbf{s}^m)_{m\geq 0},\ \mathbf{t} = (\mathbf{t}^m)_{m\geq 0} \in \prod_{m\geq 0} V^{\otimes m}.$$

Definition 3.1. *Let V be a Banach space.*[6] *We denote by $\mathbf{T}(V)$ the Banach space*

$$\mathbf{T}(V) := \left\{\mathbf{t} \in \prod_{m\geq 0} V^{\otimes m} \ :\ \|\mathbf{t}\| := \sqrt{\sum_{m\geq 0} \|\mathbf{t}^m\|^2_{V^{\otimes m}}} < \infty\right\} .$$

Define further the subset

$$\mathbf{T}_1(V) := \{\mathbf{t} \in \mathbf{T}(V) \ :\ \mathbf{t}^0 = 1\} .$$

The scaling of an element $v \in V$ by $\lambda \in \mathbb{R}$, $v \mapsto \lambda v$, extends naturally to a dilation map on $\prod_{m\geq 0} V^{\otimes m}$:

$$\delta_\lambda : \mathbf{t} \mapsto (\mathbf{t}^0, \lambda\mathbf{t}^1, \lambda^2\mathbf{t}^2, \ldots) .$$

[5] What we call strict topology above is sometimes referred to as the generalized strict topology (the term "strict topology" was previously used for the above construction on locally compact spaces; see [8]).

[6] We implicitly assume that the tensor products $V^{\otimes m}$ for all $m \geq 2$ are Banach spaces formed by the completion with respect to some admissible system of tensor norms (see [36, Sec. 3.1]).

Definition 3.2. *A tensor normalization is a continuous injective map of the form*

$$\Lambda : \mathbf{T}_1(V) \to \{\mathbf{t} \in \mathbf{T}_1(V) : \|\mathbf{t}\| \leq K\} ,$$
$$\mathbf{t} \mapsto \delta_{\lambda(\mathbf{t})}\mathbf{t} ,$$

where $K > 0$ is a constant and $\lambda : \mathbf{T}_1(V) \to (0, \infty)$ is a function.

The existence of tensor normalizations is not entirely trivial. We give a general method to construct such maps and determine their regularity properties in Appendix A; see Proposition A.2 and Corollary A.3.

In the case that $V = H$ is a Hilbert space, there exists a canonical inner product (which induces an admissible system of norms) on $H^{\otimes m}$ given on elementary tensors by

$$\langle x_1 \otimes \ldots \otimes x_m, y_1 \otimes \ldots \otimes y_m \rangle_{H^{\otimes m}} = \prod_{j=1}^{m} \langle x_j, y_j \rangle_H ,$$

and extended by linearity. Consequently, $\mathbf{T}(H)$ becomes a Hilbert space with inner product $\langle \mathbf{s}, \mathbf{t} \rangle_{\mathbf{T}(H)} = \sum_{m \geq 0} \langle \mathbf{s}^m, \mathbf{t}^m \rangle_{H^{\otimes m}}$.

4. Moments of $\mathbb{R}^d$-valued data: revisited and normalized

To motivate our construction of feature maps on pathspace, we show how composition of the monomial feature map from Example 2.4 with a tensor normalization Λ gives rise to a characteristic and universal feature map on the whole of $\mathbb{R}^d$. Note that this result is also a consequence of the classical Stone–Weierstrass and Riesz–Markov–Kakutani theorems, however these arguments usually rely on local compactness of $\mathbb{R}^d$ while our argument does not. This seems like a mere technical point at this stage, but this point plays an essential part in developing the same reasoning on pathspace (where one can not resort to local compactness arguments).

4.1. Polynomial features for $\mathbb{R}^d$.

Consider the moment map

$$\varphi : \mathbb{R}^d \to \mathbf{T}_1(\mathbb{R}^d); , \quad x \mapsto \Big(\frac{x^{\otimes m}}{m!}\Big)_{m=0}^{\infty} . \tag{6}$$

Proposition 4.1. *Let $\Lambda : \mathbf{T}_1(\mathbb{R}^d) \to \mathbf{T}_1(\mathbb{R}^d)$ be a tensor normalization. Then the map*

$$\Phi : \mathbb{R}^d \to \mathbf{T}_1(\mathbb{R}^d) , \quad \Phi = \Lambda \circ \varphi$$

(1) is a continuous injection from $\mathbb{R}^d$ into a bounded subset of $\mathbf{T}_1(\mathbb{R}^d)$,
(2) is universal to $\mathcal{F} = C_b(\mathbb{R}^d, \mathbb{R})$ equipped with the strict topology,
(3) is characteristic to $\mathcal{F}'$, the set of finite, signed Borel measures on $\mathbb{R}^d$.

Proof. (1) follows from follows the definition of a tensor normalization and the fact that $\varphi : \mathbb{R}^d \to \mathbf{T}_1(\mathbb{R}^d)$ is a continuous injection. To show (2), we claim that the family of functions

$$\{x \mapsto \langle \ell, \Phi(x) \rangle, \ell \in (\mathbb{R}^m)' , \ m \geq 0\} \subset C_b(\mathbb{R}^d, \mathbb{R}) \tag{7}$$

satisfies the conditions of item (2) of Theorem 2.6. Indeed, for $m, n \geq 0$ and $\ell \in \left(\mathbb{R}^d\right)^{\otimes m}$, $\tilde{\ell} \in \left(\mathbb{R}^d\right)^{\otimes n}$ we have

$$\langle \ell, \Phi(x) \rangle \langle \tilde{\ell}, \Phi(x) \rangle = \lambda(\varphi(x))^{m+n} \langle \ell, \frac{x^{\otimes m}}{m!} \rangle \langle \tilde{\ell}, \frac{x^{\otimes n}}{n!} \rangle = \binom{m+n}{m} \langle \ell \otimes \tilde{\ell}, \Phi(x) \rangle$$

which shows that (7) is a subalgebra of $C_b(\mathbb{R}^d, \mathbb{R})$. The family (7) separates points since Φ is injective. Finally, note that $\Phi(x) \neq 0$ for all $x \in \mathbb{R}^d$. It follows by item (2) of Theorem 2.6 that Φ is universal to $C_b(\mathbb{R}^d, \mathbb{R})$ equipped with the strict topology which shows (2). Point (3) now follows by item (3) of Theorem 2.6. □

4.2. Kernelization.

For applications, we are usually only interested in moments up to a finite degree $M \geq 1$ (e.g. due to computational restrictions or to avoid overfitting in learning algorithms). Therefore we define

$$\varphi_M(x) = \Big(1, x, \frac{x^{\otimes 2}}{2!}, \ldots, \frac{x^{\otimes M}}{M!}, 0, 0, 0, \ldots\Big) \in \mathbf{T}(\mathbb{R}^d)$$

and analogously define the normalized monomials $\Phi_M = \Lambda \circ \varphi_M : \mathbb{R}^d \to \mathbf{T}_1(\mathbb{R}^d)$. The number of monomials in d-variables of degree M equals $\binom{d+M}{d}$, hence the number of coordinates in φ_M grows rapidly. An important insight from statistical learning is that many learning algorithms only rely on inner products $\langle \Phi(x), \Phi(y) \rangle$, and that such inner products can be much more efficient to evaluate than Φ itself ("the kernel trick"). For d-dimensional variables, such kernel tricks are classical, but as a precursor to our feature kernelization on path space, we spell out the details. We refer to [18, 44] for more background on kernel learning.

Proposition 4.2. *Let $\psi : [1,\infty) \to [1,\infty)$ and $\Lambda : \mathbf{T}_1(\mathbb{R}^d) \to \mathbf{T}_1(\mathbb{R}^d)$ be as in Proposition A.2 (not necessarily a tensor normalization). Then for any $M \geq 1$, the inner product*

$$k_M : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad (x,y) \mapsto \langle \Phi_M(x), \Phi_M(y) \rangle$$

can be evaluated in $O(d+M+q)$ computational steps, where q is the combined cost of a single evaluation of ψ and of finding the unique non-negative root of a polynomial $P(\lambda) = \sum_{m=0}^{M} a_m \lambda^{2m}$ with $a_0 \leq 0 \leq a_1, \dots, a_M$.

Proof. For $x, y \in \mathbb{R}^d$, observe that the sequence $(\langle x,y\rangle^{2m})_{m=0}^M$ can be evaluated in $O(d+M)$ operations: $O(d)$ to evaluate $\langle x,y\rangle$, and a further $O(M)$ to evaluate the powers $\langle x,y\rangle^{2m}$. In particular, the kernel

$$\kappa_M(x,y) := \langle \varphi_M(x), \varphi_M(y) \rangle ,$$

and the sequence $(\|x^{\otimes m}\|^2)_{m=0}^M = (\langle x,x\rangle^{2m})_{m=0}^M$ can be evaluated in $O(d+M)$.

Observe that $\lambda(\varphi_M(x)) \geq 0$ is the unique non-negative root of the polynomial

$$P(\lambda) := \|\delta_\lambda \varphi_M(x)\|^2 - \psi(\|\varphi_M(x)\|) = \Big(\sum_{m=0}^{M} \lambda^{2m} \|x^{\otimes m}\|^2 / m! \Big) - \psi(\|\varphi_M(x)\|) ,$$

which can be found in $O(q)$ by assumption once the sequence $(\|x^{\otimes m}\|^2\|)_{m=0}^M$ is given. The conclusion since $k_M(x,y) = \kappa_M(\lambda(\varphi_M(x))x, \lambda(\varphi_M(y))y)$. □

Remark 4.3. *Compared to $O(d^M)$ which is the complexity of evaluating $\Phi_M(x)$ and then computing the inner product, the above method gives an exponential reduction in cost and allows for learning in high dimensions. In fact, the above is different from the usual Veronese embedding $(x,y) \mapsto (1 + \langle x,y\rangle_{\mathbb{R}^d})^M$, and is better suited for the generalization to monomials on path space, see* [33].

Remark 4.4. *The cost to evaluate ψ and find the non-negative root of $P(\lambda)$ is typically negligible. See Remark 8.1 for implementations in our experiments.*

Remark 4.5 (Moments of data in non-linear spaces)**.** *The above immediately extends from data in $\mathbb{R}^d$ to general topological spaces $\mathcal{X}$ by replacing $\langle \cdot,\cdot\rangle_{\mathbb{R}^d}$ with a kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. This turns κ into a kernel k_M on $\mathcal{X}$ that has the (normalized moments) of the feature map $\kappa_x := \kappa(x,\cdot) : \mathcal{X} \to \mathbb{R}$ as features.*

5. Features for path-valued data: the bounded variation case

Throughout this section, let V be a Banach space. We combine normalization with the so-called signature map to construct a universal feature map

$$\overline{\Phi} : C^1 \to \mathbf{T}_1(V) ,$$

where C^1 denotes the subset of $C([0,1],V)$ of bounded variation paths. In fact, we first construct another feature map

$$\Phi : \mathcal{P}^1 \to \mathbf{T}_1(V) ,$$

where $\mathcal{P}^1 := C^1/\sim_t$ is the quotient space under so-called "tree-like equivalence" $x \sim_t y$. This is often a natural equivalence relation on pathspace [30], since it factors out natural transformations of paths, e.g. if x is a time-change of y then $x \sim_t y$.

5.1. Signatures: monomials of paths.

An insight that played a prominent role in stochastic analysis is that the map

$$\mathrm{S} : C^1 \to \mathbf{T}_1(V) , \quad x \mapsto \Big(1, \int dx, \int dx^{\otimes 2}, \int dx^{\otimes 3}, \dots\Big) \tag{8}$$

is the natural generalization of the monomial feature map (6) to pathspace. We consider first paths of bounded variation with starting point zero

$$C^1 := \{x \in C([0,1],V) : \|x\|_{1\text{-var}} < \infty ,\ x(0) = 0\} , \quad \|x\|_{1\text{-var}} := \sup_{\substack{n\geq 1 \\ 0\leq t_1 \leq \dots \leq t_n \leq 1}} \sum_{i=1}^{n-1} \|x(t_{i+1}) - x(t_i)\| , \tag{9}$$

since this allows us to use Riemann–Stieltjes integrals to define

$$\int dx^{\otimes m} := \int_{0\leq t_1 \leq \dots \leq t_m \leq 1} dx(t_1) \otimes \dots \otimes dx(t_m) \in V^{\otimes m} . \tag{10}$$

The general unbounded variation/rough path case is covered in Section 6.

Example 5.1. *If $V = \mathbb{R}^d$, $\int dx^{\otimes m} = \Big(\int_{0\leq t_1 \leq \dots \leq t_m \leq 1} dx^{i_1}(t_1) \cdots dx^{i_m}(t_m)\Big)_{i_1,\dots,i_m \in \{1,\dots,d\}}$ where $x^i(t)$ denotes the i-th coordinate of $x(t)$.*

Example 5.2. *Let $V = \mathbb{R}^d$. Consider the linear path $x(t) = t.v$ for a fixed vector $v \in \mathbb{R}^d$. A simple calculation shows* $\mathrm{S}(x) = (\frac{v^{\otimes m}}{m!})$. *This shows that* S *is a generalization of the monomial feature map on $\mathbb{R}^d$.*

An algebraic reason why (10) can be seen as sequence of monomials on path space C^1, is that their linear span is closed under multiplication. The so-called shuffle product makes this precise: considering for simplicity $V = \mathbb{R}^d$, we have for $\mathbf{i} = (i_1, \dots, i_m) \in \{1, \dots, d\}^m, \mathbf{j} = (j_1, \dots, j_n) \in \{1, \dots, d\}^n$

$$\int dx_{\mathbf{i}}^{\otimes m} \int dx_{\mathbf{j}}^{\otimes n} = \sum_{\mathbf{k}} \int dx_{\mathbf{k}}^{\otimes (m+n)} , \tag{11}$$

where the sum is taken over all $\mathbf{k} = (k_1, \dots, k_{m+n})$ that are shuffles[7] of $\mathbf{i}$ and $\mathbf{j}$, see [37, Thm. 2.15]. The signature map is furthermore injective up to tree-like equivalence.

Theorem 5.3 ([4])**.** *The map $\mathrm{S} : C^1 \to \mathbf{T}_1(V)$ is injective up to tree-like equivalence: for $x, y \in C^1$, $\mathrm{S}(x) = \mathrm{S}(y)$ if and only if $x \sim_t y$ where $\sim_t$ denotes tree-like equivalence.*

We recall the definition of tree-like equivalence in appendix B.

Example 5.4. *An important case of tree-like equivalence is when y is a reparameterization of x i.e. there exists an continuous increasing bijection $\tau : [0, 1] \to [0, 1]$ such that $y = x \circ \tau$.*

5.2. **Features for unparameterized paths.** The identification of tree-like equivalent paths can be an extremely powerful dimensionality reduction for many learning tasks as seen by the success of dynamic time warping and Fréchet distances. This motivates the following definition.

Definition 5.5. *We define the space of unparameterised bounded variation paths $\mathcal{P}^1$ as the set of equivalence classes $C^1 / \sim_t$ equipped with the quotient topology.*

In complete analogy to our toy example of the monomial feature map for $\mathbb{R}^d$ in Proposition 4.1, we normalize the "ordered moments map" S to get a universal and characteristic feature map.

Theorem 5.6. *Let $\Lambda : \mathbf{T}_1(V) \to \mathbf{T}_1(V)$ be a tensor normalization. The normalized signature*

$$\Phi : \mathcal{P}^1 \to \mathbf{T}_1(V), \quad \Phi = \Lambda \circ \mathrm{S}$$

(1) is a continuous injection from $\mathcal{P}^1$ into a bounded subset of $\mathbf{T}_1(V)$,
(2) is universal to $\mathcal{F} := C_b(\mathcal{P}^1, \mathbb{R})$ equipped with the strict topology,
(3) is characteristic to the space of finite regular Borel measures on $\mathcal{P}^1$.

Proof. (1) follows from the definition of tensor normalization and the fact that $\mathrm{S} : \mathcal{P}^1 \to \mathbf{T}_1(V)$ is continuous [36, Thm. 3.1.3] and injective by Theorem 5.3. For (2), define $L := \bigoplus_{m \geq 0} (V^{\otimes m})'$, which we identify with a dense subspace of $\mathbf{T}(V)'$ via $\ell(\mathbf{t}) = \sum_{m \geq 0} \langle \ell^m, \mathbf{t}^m \rangle$. Define further $\mathcal{F}_0 := \{\ell \circ \Phi : \ell \in L\} \subset \mathcal{F}$. By point (1), $\mathcal{F}_0$ separates the points $\mathcal{P}^1$. Furthermore, S takes values in the group-like elements of $\mathbf{T}(V)$ [10, Cor. 3.9] which implies $\mathcal{F}_0$ is closed under multiplication (in the case $V = \mathbb{R}^d$, this is equivalent to the shuffle product (11)). It follows that $\mathcal{F}_0$ satisfies the assumptions of item (2) of Theorem 2.6. Hence $\mathcal{F}_0$ is dense in $\mathcal{F}$ under the strict topology, which proves (2). Point (3) in turn follows from Theorem 2.3 and item (3) of Theorem 2.6. □

5.3. **Features for parameterized paths.** We now derive a feature map $\overline{\Phi}$ from Φ, that also distinguishes between tree-like equivalent paths. This is useful for situations where time-parameterization matters (e.g. financial data). $\overline{\Phi}$ is constructed by simply adding time as a coordinate to the path before computing Φ. Denote with $\overline{V} := V \oplus \mathbb{R}$ the direct sum of the Banach spaces V and $\mathbb{R}$ and map

$$C^1(V) \ni x \mapsto \overline{x} \in C^1(\overline{V}) \text{ where } \overline{x}(t) := (x(t), t) .$$

Corollary 5.7. *Let $\Lambda : \mathbf{T}_1(\overline{V}) \to \mathbf{T}_1(\overline{V})$ be a tensor normalization. The normalized signature*

$$\overline{\Phi} : C^1 \to \mathbf{T}_1(\overline{V}) , \quad x \mapsto \Lambda \circ \mathrm{S}(\overline{x})$$

(1) is a continuous injection from C^1 into a bounded subset of $\mathbf{T}_1(\overline{V})$,
(2) is universal to $C_b(C^1, \mathbb{R})$ equipped with the strict topology,
(3) is characteristic to the space of finite regular Borel measures on C^1.

Proof. The map $x \mapsto \overline{x}$ is continuous and for $x, y \in C^1$, it holds by strict monotonicity of the second component that $\overline{x} \sim_t \overline{y}$ if and only if $x = y$. The claim then follows from Theorem 5.6. □

[7] A shuffle $\mathbf{k} = (k_1, \dots, k_{m+n})$ of $\mathbf{i}$ and $\mathbf{j}$ is a permutations of $(i_1, \dots, i_m, j_1, \dots, j_n)$ subject to the condition that the order of elements in $\mathbf{i}$ and $\mathbf{j}$ is preserved.

5.4. **Characterizing laws of stochastic processes.** For applications in statistics, it is sufficient to consider Borel probability measures and the previous results reduce to the following corollary.

Corollary 5.8. *Let (Ω, Σ) be a measurable space and $X = (X_t)_{t\in[0,1]}$ a stochastic process defined on (Ω, Σ). Let $\mathbb{P}, \mathbb{Q}$ two probability measures such that $X(\omega) \in C^1$ almost surely (under both $\mathbb{P}$ and $\mathbb{Q}$). Then*

$$\mathbb{E}_{\mathbb{P}}[\overline{\Phi}(X)] = \mathbb{E}_{\mathbb{Q}}[\overline{\Phi}(X)] \quad \textit{iff} \quad \mathbb{P} = \mathbb{Q}.$$

Remark 5.9.

- *Under additional integrability assumptions, a similar result holds for the unnormalized moments $\mathbb{E}[\mathrm{S}(X)]$, see* [15, Prop. 6.1]. *This is an analogue of the classical moment problem, the proof of which, due to the unboundedness of the map* S*, relies on a non-commutative Fourier transform.*
- *The tensor normalization $\Lambda = \delta_{\lambda(\cdot)}\cdot$ can be seen as a rescaling of the path x, and the latter is equivalent to scalar multiplication: for $\lambda \in \mathbb{R}$ and a path $x \in C^1$, set $x^\lambda(t) = \lambda.x(t)$. Then $\mathrm{S}(x^\lambda) = \delta_\lambda \mathrm{S}(x)$, so that $\Phi(x) = \delta_{\lambda(\mathrm{S}(x))}\,\mathrm{S}(x) = \mathrm{S}(x^{\lambda(\mathrm{S}(x))})$. Thus, in contrast to a Fourier approach which deals with unboundedness by spinning around unitary groups, our approach injects moments into a ball.*

5.5. **Lifting state space features to path space features.** Any map $\varphi : \mathcal{X} \to V$ naturally lifts a path $x : [0, 1] \to \mathcal{X}$ to a path $\varphi(x) = \varphi \circ x : [0, 1] \to V$ that evolves in the linear space V. Thus it is natural to consider $\mathrm{S} \circ \varphi$ as feature map for $C([0, 1], \mathcal{X})$. This allows to learn from paths that evolve in a general topological space $\mathcal{X}$ that does not carry a linear structure — at least not a canonical linear structure — provided that a feature map $\varphi : \mathcal{X} \to V$ is given.

Example 5.10. *Let $\mathcal{X}$ be the space of graphs. A possible feature map φ is to map a graph to its adjacency matrix. Other choices are possible, such as those arising from kernels as described in Section 7.1.*

Even for paths in a linear space $\mathcal{X} = \mathbb{R}^d$, this can have huge benefits in terms of efficiency.

Example 5.11. *Consider a collection of points $\{v_1, \dots, v_n\} \subset \mathbb{R}^d$. To learn non-linear relations from this data, one could consider basic monomials $(1, v_i, \frac{v_i^{\otimes 2}}{2!}, \frac{v_i^{\otimes 3}}{3!}, \dots)$, which is just $\mathrm{S}(x_i)$ for the linear path $x_i(t) = v_i \cdot t$. However, there is no reason to prefer this as a basis for polynomials to Tchebyscheff, Legendre, Hermite polynomials, etc., or more generally other non-linear functions such as wavelets or Fourier series. In fact, it may be much more efficient to find a feature map $\varphi : \mathbb{R}^d \to E$ into some high-dimensional space and consider the monomials $(1, \varphi(v_i), \frac{\varphi(v_i)^{\otimes 2}}{2!}, \frac{\varphi(v_i)^{\otimes 3}}{3!}, \dots)$. Depending on $\{v_1, \dots, v_n\}$, different choices of φ can lead to a much more efficient learning.*

Definition 5.12. *Let $\varphi : \mathcal{X} \to V$ be a feature map and $\Lambda : \mathbf{T}_1(V) \to \mathbf{T}_1(V)$ a tensor normalization. Denote with $C^1(\mathcal{X})$ the subset of paths x in $C([0, 1], \mathcal{X})$ such that $\varphi(x) \in C^1$. We call the map*

$$\Phi^\varphi : C^1(\mathcal{X}) \to \mathbf{T}_1(V), \quad \Phi^\varphi := \Lambda \circ \mathrm{S} \circ \varphi$$

the normalized signature lift of the feature map φ.

Proposition 5.13. *Let notation be as in Definition 5.12. Suppose $\varphi : \mathcal{X} \to V$ is injective. Let $x_0 \in \mathcal{X}$ and $C^1_{x_0}(\mathcal{X}) = \{x \in C^1(\mathcal{X}) : x(0) = x_0\}$ equipped with initial topology induced by $\varphi : C^1_{x_0}(\mathcal{X}) \to C^1$. Denote the quotient topological space under tree-like equivalence $\mathcal{P}^1_{x_0}(\mathcal{X}) = C^1_{x_0}(\mathcal{X})/\sim_t$. Then Φ^φ*

(1) is a continuous injection from $\mathcal{P}^1_{x_0}(\mathcal{X})$ into a bounded subset of $\mathbf{T}_1(V)$,
(2) is universal to $C_b(\mathcal{P}^1_{x_0}(\mathcal{X}), \mathbb{R})$ equipped with the strict topology,
(3) is characteristic to the space of finite regular Borel measures on $\mathcal{P}^1_{x_0}(\mathcal{X})$.

The same statement holds when $\mathcal{P}^1_{x_0}(\mathcal{X})$ is replaced by $C^1_{x_0}(\mathcal{X})$ and Φ^φ is replaced by $\Lambda \circ \mathrm{S}(\overline{\varphi(x)})$.

Proof. Injectivity of φ implies that for all $x, y \in C^1_{x_0}(\mathcal{X})$, $\varphi(x) \sim_t \varphi(y)$ if and only if $x \sim_t y$. All the desired claims now follow from Theorem 5.6 and Corollary 5.7. □

Remark 5.14. *If $\mathcal{X} = V$ and $\varphi = \mathrm{id} : \mathcal{X} \to V$, we recover our usual feature map $\Phi^{\mathrm{id}} = \Phi$ for $\mathcal{P}^1$.*

For sequence-valued data (the set of sequences in $\mathcal{X}$ is locally compact, in contrast to the set of paths in $\mathcal{X}$), this was introduced in [33] in the context of kernel learning and called "sequentialization". Already for sequences in $\mathcal{X} = \mathbb{R}^d$ such an approach turned out be very powerful and nonlinear choices for $\varphi : \mathbb{R}^d \to V$ generically beat $\varphi = \mathrm{id}$ in the context of supervised sequence classification. We will see in Section 8 that the same holds for the metric on laws of stochastic processes and hypothesis testing.

Remark 5.15. *Example 5.11 motivated the use of a non-linear φ by arguing that a special case of learning from paths is learning from vector-valued data and for the latter the use of non-linear feature maps is classic. On a more formal level,*

$$\mathrm{S}(\varphi(x)) = \int d\varphi(x)^{\otimes m} = \int D^{\otimes m}_\varphi(x) dx^{\otimes m}$$

if D_φ denotes the Jacobian of φ. Thus the appearance of the Jacobian distorts the integrals and has the potential to capture the properties of the observed data better.

6. Features for path-valued data: the rough path case

We enlarge the domain of the feature map Φ [resp. $\overline{\Phi}$] to paths of unbounded 1-variation. This allows to include examples of stochastic processes such as semimartingales, Markov processes, and Gaussian processes. The difficulty is that the iterated integrals $\int dx^{\otimes m}$ can no longer be defined by Riemann–Stieltjes integration. Rough paths theory provides a complete integration theory for a large class of processes (see Example C.3); for the special case of (semi)martingales, the theory agrees with Itô integration. We present the main ideas in this section and give a detailed account in Appendix C. We also present how the ideas from this section generalise to branched rough paths in Appendix D.

Throughout this section, let V be a Banach space. Rough path theory provides a family of spaces $(C^p)_{p\geq 1}$ called geometric p-rough paths over V. These spaces satisfy the inclusions $C^p \subset C^q$ for $p \leq q$ (in much the same way as $\ell^p \subset \ell^q$ for the family $(\ell^p)_{p\geq 1}$), and larger values of p allow for paths with "rougher" trajectories. For every $p \geq 1$, there exists a map

$$\mathrm{S} : C^p \to \mathbf{T}_1(V)$$

that has the same properties as the signature map from Section 5. In fact, for $p = 1$, C^1 is just the space of bounded 1-variation paths and we recover the setting of Section 5. Similar to before, we define the quotient space $\mathcal{P}^p = C^p/\sim_t$ where $\sim_t$ denotes tree-like equivalence. The following result generalises Theorem 5.6 and Corollary 5.7, and is proved in the same way using Theorem C.5.

Theorem 6.1. *Let $p \geq 1$ and $\Lambda : \mathbf{T}_1(V) \to \mathbf{T}_1(V)$ a tensor normalization. The map*

$$\Phi : \mathcal{P}^p \to \mathbf{T}_1(V)\,, \quad \Phi = \Lambda \circ \mathrm{S}$$

(1) is a continuous injection from $\mathcal{P}^p$ into a bounded subset of $\mathbf{T}_1(V)$,
(2) is universal to $\mathcal{F} := C_b(\mathcal{P}^p, \mathbb{R})$ equipped with the strict topology,
(3) is characteristic to the space of finite regular Borel measures on $\mathcal{P}^p$.

The same statement holds when $\mathcal{P}^p$ is replaced by C^p and Φ is replaced by $\overline{\Phi}(\mathbf{x}) := \Lambda \circ \mathrm{S}(\overline{\mathbf{x}})$, see Appendix C.2.

Corollary 6.2. *Let μ_1, μ_2 be finite Borel measures on $C([0,1], V)$. Suppose there exist $p \geq 1$ and measurable maps $S_1, S_2 : C([0,1], V) \to C^p$ such that for $i = 1, 2$, $S_i(x)$ is a rough path lift of x for μ_i-a.e. $x \in C([0,1], \mathbb{R}^d)$. Then*

$$\mu_1 = \mu_2 \quad \text{iff} \quad \mu_1[\overline{\Phi}(\cdot)^m] = \mu_2[\overline{\Phi}(\cdot)^m] \quad \forall m \geq 1\,.$$

As discussed in Example C.3, μ_i can be chosen to be probability measures arising from a wide range of processes, including many Gaussian and Markov processes, and all continuous semimartingales.

7. A computable metric for laws of stochastic processes

A natural distance for probability measures on a space $\mathcal{X}$ is to fix a class of functions $\mathcal{G} \subset \mathbb{R}^{\mathcal{X}}$ and define the "maximum mean distance" (MMD) as

$$d(\mu, \nu) = \sup_{f \in \mathcal{G}} |\mathbb{E}_{X\sim\mu}[f(X)] - \mathbb{E}_{Y\sim\nu}[f(Y)]|$$

If $\mathcal{G}$ is sufficiently large, then d defines a metric between probability measures, see [42, 39]. However, the supremum makes it hard to compute or estimate $d(\mu, \nu)$. An insight from kernel learning is that if $\mathcal{G}$ is a unit ball of a RKHS $(\mathcal{H}, \mathrm{k})$, then the reproducing property implies

$$d(\mu, \nu) = \mathbb{E}[\mathrm{k}(X, X')] - 2\mathbb{E}[\mathrm{k}(X, Y)] + \mathbb{E}[\mathrm{k}(Y, Y')]\,.$$

The above can be simply estimated from finite samples of μ and ν provided k is cheap to evaluate.

In this section we discuss the case when $\mathcal{X} = C^p$ is the space of p-rough paths and μ and ν are the laws of stochastic processes $X = (X_t)$ and $Y = (Y_t)$. We use the feature map Φ to define a kernel on pathspace and discuss properties of the MMD it induces.

7.1. Kernel learning.

We briefly recall notation and terminology from kernel learning [18, 44, 2]. Throughout this section, let $\mathcal{X}$ be a topological space. Let us fix a feature map $\mathcal{X} \to \mathbb{R}^{\mathcal{X}}$, $x \mapsto \mathrm{k}_x$. Under the condition that $\mathrm{k}(x, y) := \mathrm{k}_x(y)$ is a symmetric positive definite kernel (which we assume henceforth), the completion of $\mathcal{H}_0 := \{\mathrm{k}_x : x \in \mathcal{X}\}$ under the inner product $\langle \mathrm{k}_x, \mathrm{k}_y \rangle = \mathrm{k}(x, y)$ is the RKHS associated with k. Let us further fix a locally convex TVS $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ containing $\mathcal{H}_0$ for which the inclusion map $\mathcal{H}_0 \hookrightarrow \mathcal{F}$ is continuous.

Definition 7.1 ([46])**.** *We say that the kernel* k *is*

- *universal to $\mathcal{F}$ if the kernel embedding $\iota = \mathrm{id} : \mathcal{H}_0 \hookrightarrow \mathcal{F},\ f \mapsto f$, is dense,*
- *characteristic to $\mathcal{F}'$ if the kernel mean embedding $\mu : \mathcal{F}' \to \mathcal{H}_0',\ D \mapsto D|_{\mathcal{H}_0}$ is injective.*

Proposition 7.2 ([46])**.** *It holds that*

- k *is universal to* $\mathcal{F}$ *iff* k *is characteristic to* $\mathcal{F}'$.
- *the kernel embedding* ι *and the kernel mean embedding* μ *are transpose*

$$\iota^\star = \mu \text{ and } \mu^\star = \iota \ .$$

A common way to construct the feature map $x \mapsto \mathrm{k}_x$ is through another feature map $\Phi : \mathcal{X} \to E$ taking values in an inner product space $(E, \langle\cdot,\cdot\rangle)$. Then one defines $\mathrm{k}_x(y) := \mathrm{k}(x,y) := \langle\Phi(x), \Phi(y)\rangle$. As expected, universality and characteristicness of Φ and k are equivalent.

Proposition 7.3. *Consider a map* $\Phi : \mathcal{X} \to E$ *into an inner product space* $(E, \langle\cdot,\cdot\rangle)$ *such that* $\mathrm{k}(x,y) = \langle\Phi(x), \Phi(y)\rangle$. *Then*

- Φ *is universal to* $\mathcal{F}$ *iff* k *is universal to* $\mathcal{F}$,
- *is characteristic to* $\mathcal{F}'$ *iff* k *is characteristic to* $\mathcal{F}'$.

Proof. See Proposition E.3. □

7.2. **Maximum mean distances.**

Definition 7.4. *Let* $\mathcal{F}$ *be a TVS. For a subset* $\mathcal{G} \subset \mathcal{F}$, *we call*

$$d_{\mathcal{G}} : \mathcal{F}' \times \mathcal{F}' \to [0,\infty), \quad d_{\mathcal{G}}(D_1, D_2) := \sup_{f\in\mathcal{G}} |D_1(f) - D_2(f)|$$

the maximum mean distance (MMD) over $\mathcal{G}$.

The function $d_{\mathcal{G}}$ is a pseudo-metric. It is furthermore a metric if and only if $\mathcal{G}$ is large enough to separate the points of $\mathcal{F}'$. The most popular form in which the above metric appears is when $\mathcal{F}$ carries a (semi-)norm, $\mathcal{G}$ is a unit ball in this (semi-)norm, and $d_{\mathcal{G}}$ is restricted to probability measures.

Example 7.5. *Let* $\mathcal{X} = \mathbb{R}^d$ *and* $\mathcal{F} = C(\mathcal{X}, \mathbb{R})$. *Then the supremum norm gives the total variation metric; the Lipschitz norm gives the Wasserstein distance; the sum of Lipschitz and supremum norms gives the Dudley metric.*

A well-known disadvantage of an MMD is that it metrizes a topology that is usually not comparable (i.e. neither weaker nor stronger) to weak convergence; [42, 39]. Progress on this question was made more recently: [46, Thm. 55] characterises kernels for which the associated MMD metrizes the topology of weak convergence[8] on the space of probability measures over a locally compact space $\mathcal{X}$. The question of whether the same characterization holds for any Polish space $\mathcal{X}$ was left open in [46]. The following results shows that this is not the case.

Proposition 7.6 (Local compactness necessary in Theorem 55 in [46])**.** *There exits a Polish non-locally compact space* $\mathcal{X}$, *a closed subspace* $\mathcal{F} \subset C_b(\mathcal{X}, \mathbb{C})$ *(equipped with the uniform topology), a Hilbert space* E, *and a map* $\Phi : \mathcal{X} \to E$ *with the following properties:*

(1) the positive definite kernel $\mathrm{k}(x,y) := \langle\Phi(x), \Phi(y)\rangle_E$ *is bounded and continuous, and* $\mathcal{H}_0$ *embeds continuously and densely into* $\mathcal{F}$ *(so in particular,* k *is characteristic to* $\mathcal{F}'$*),*

(2) $\mathcal{F}'$ *contains the probability measures* $\mathcal{P}$ *on* $\mathcal{X}$,

(3) the metric

$$d_{\mathrm{k}}(D_1, D_2) := \sup_{f\in\mathcal{H}, \|f\|_{\mathcal{H}}\le 1} |D_1(f) - D_2(f)| \ ,$$

where $\mathcal{H}$ *is the completion of* $\mathcal{H}_0$, *does not metrize the weak topology on* $\mathcal{P}$.

Proposition 7.6 is shown by an explicit construction of a counterexample; see item (4) of Theorem 7.7 for another counterexample which arises naturally from the normalized signature map Φ.

Proof of Proposition 7.6. Let $\mathcal{X}$ be a separable infinite-dimensional Hilbert space with an orthonormal basis $x_1, x_2, \ldots$. Consider the countable set $Q \subset \mathbb{Q}^{\mathbb{N}}$ of sequences of rational numbers $q = (q_1, q_2, \ldots)$ which eventually vanish. Let $w : Q \to (0,1)$ be a function for which $\sum_{q\in Q} w(q)^2 = 1$. For every $q \in Q$, consider the function

$$f_q : \mathcal{X} \to \mathbb{C}, \quad f_q(x) := w(q) \exp\Big(i\Big\langle \sum_{n\ge 1} q_n x_n, x\Big\rangle_{\mathcal{X}}\Big)$$

and note that f_q is a continuous bounded function on $\mathcal{X}$ for every $q \in Q$. Consider the span $\mathcal{F}_0 = \mathrm{span}\{f_q \mid q \in Q\} \subset C_b(\mathcal{X}, \mathbb{C})$ and let $\mathcal{F}$ denote the closure in $C_b(\mathcal{X}, \mathbb{C})$ of $\mathcal{F}_0$ (under the uniform norm). Define the Hilbert space $E = \ell^2(Q)$ as the space of square-summable functions $f : Q \to \mathbb{C}$ with the inner product $\langle f, g\rangle_E := \sum_{q\in Q} f(q)\overline{g(q)}$, and consider the map

$$\Phi : \mathcal{X} \to E, \quad \Phi(x)(q) := f_q(x) \ .$$

As shown in Appendix F, Φ fulfils points (1), (2), and (3). □

[8] also called the narrow topology

7.3. **The normalized signature kernel and its MMD.** The feature map Φ is universal and characteristic for unparameterized paths that evolve in a Banach space V. For paths that evolve in a Hilbert space $(H, \langle\cdot,\cdot\rangle)$, Φ takes values in the Hilbert space $\mathbf{T}(H)$ (see Section 3). This immediately gives rise to a kernel on $\mathcal{P}^p$ that is also universal and characteristic.

Theorem 7.7. *Let $p \geq 1$ and $\Lambda : \mathbf{T}_1(H) \to \mathbf{T}_1(H)$ a tensor normalization. Define*

$$\mathrm{k} : \mathcal{P}^p \times \mathcal{P}^p \to \mathbb{R}, \quad (x, y) \mapsto \langle \Phi(x), \Phi(y) \rangle$$

where $\Phi = \Lambda \circ \mathrm{S}$ is the normalized signature map. We call k *the normalized signature kernel and denote with*

$$d_{\mathrm{k}}(\mu, \nu) := \sup_{f \in \mathbf{T}(H) : \|f\| \leq 1} \left| \int f(x)\mu(dx) - \int f(y)\nu(dy) \right|$$

the associated MMD. Then k *is a bounded, continuous, positive definite kernel and*

(1) k *is universal to $C_b(\mathcal{P}^p, \mathbb{R})$ equipped with the strict topology,*
(2) k *is characteristic to finite, signed Borel measures on $\mathcal{P}^p$,*
(3) d_{k} is a metric on the space of finite, signed Borel measures on $\mathcal{P}^p$, and

$$d_{\mathrm{k}}(\mu, \nu) = \mathbb{E}[\mathrm{k}(X, X')] - 2\mathbb{E}[\mathrm{k}(X, Y)] + \mathbb{E}[\mathrm{k}(Y, Y')]$$

(4) for probability measures on $\mathcal{P}^p$, convergence in d_{k} does not imply weak convergence. If H is finite dimensional, then weak convergence implies convergence in d_{k}.

The same statement holds when $\mathcal{P}^p$ is replaced by $\mathcal{C}^p$ and Φ in $\mathrm{k}(x, y) = \langle \Phi(x), \Phi(x) \rangle$ is replaced by $\overline{\Phi}(\mathbf{x}) := \Lambda \circ \mathrm{S}(\overline{\mathbf{x}})$.

Proof. Point (4) follows from Proposition C.8, and all remaining claims follow from combining Theorem 6.1 and Proposition 7.3. □

7.4. **Lifting state space kernels to path space kernels.** In Section 5.5 we pointed out, that our construction extends to paths $x \in C([0, 1], \mathcal{X})$ that evolve in a topological space if a feature map $\varphi : \mathcal{X} \to V$ for the state space $\mathcal{X}$ is provided. Namely, $\Phi^\varphi = \Phi \circ \varphi : \mathcal{P}^1(\mathcal{X}) \to \mathbf{T}(V)$ becomes a universal and characteristic feature map for paths that evolve in $\mathcal{X}$.

The same remark applies to the normalized signature kernel: suppose $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel on $\mathcal{X}$ with RKHS H. Let $\varphi : \mathcal{X} \to H$, $\varphi(x) := \kappa(x, \cdot)$ be the natural feature map. For $x \in C([0, 1], \mathcal{X})$ denote by $x \mapsto \kappa_x := (\varphi_{x(t)})_t \in C([0, 1], H)$ the lift of x to a path evolving in H. Letting k denote the kernel in Theorem 7.7, it holds that

$$(x, y) \mapsto \mathrm{k}^\kappa(x, y) := \mathrm{k}(\kappa_x, \kappa_y)$$

is a kernel for paths that evolve in general topological spaces $\mathcal{X}$.

Proposition 7.8. *Let (H, κ) be a RKHS on $\mathcal{X}$ and $\varphi : \mathcal{X} \to H$, $\varphi(x) := \kappa(x, \cdot)$. Using the notation from Proposition 5.13, we call*

$$\mathrm{k}^\kappa : \mathcal{P}^1(\mathcal{X}) \times \mathcal{P}^1(\mathcal{X}) \to \mathbb{R}, \quad (x, y) \mapsto \mathrm{k}(\kappa_x, \kappa_y)$$

the normalized signature lift of the kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then k^κ is a bounded, continuous, positive definite kernel and for any $x_0 \in \mathcal{X}$

(1) k^κ is universal to $C_b(\mathcal{P}^1_{x_0}(\mathcal{X}), \mathbb{R})$ equipped with the strict topology,
(2) k^κ is characteristic to finite, signed Borel measures on $\mathcal{P}^1_{x_0}(\mathcal{X})$,
(3) d_{k^κ} is a metric on the space of finite, signed Borel measures on $\mathcal{P}^1_{x_0}(\mathcal{X})$.

The same statement holds when $\mathcal{P}^1_{x_0}(\mathcal{X})$ is replaced by $C^1_{x_0}(\mathcal{X})$ and Φ (implicit in the definition of k*) is replaced by $\overline{\Phi}(x) := \Lambda \circ \mathrm{S}(\overline{x})$.*

Remark 7.9. *If $\mathcal{X}$ is itself a Hilbert space, the choice $\kappa(\cdot,\cdot) = \langle\cdot,\cdot\rangle$ corresponds to $\varphi = \mathrm{id}$ and recovers the signature kernel* k *from Theorem 7.7,* $\mathrm{k}^{\langle\cdot,\cdot\rangle} = \mathrm{k}$.

Following the discussion at the end of Section 5.5, our experiments in Section 8 show that even if the original state space is $\mathcal{X} = \mathbb{R}^d$, the composition with a non-linear kernel κ that lifts paths evolving in $\mathbb{R}^d$ to paths evolving in an infinite dimensional Hilbert space can be very useful.

7.5. **Computing the signature kernel.** So far we have ignored the computation cost of evaluating the normalized signature map Φ and the associated kernel k. Firstly, we wish to truncate the signature after a given tensor degree M to avoid overfitting in learning tasks and due to computational restrictions — recall that the signature is a generalization of monomials and for the latter this truncation is classical. That is, we denote with $\mathrm{S}_{\leq M}(x) = \left(1, \int dx, \ldots, \int dx^{\otimes M}, 0, 0, \ldots\right)$ the truncated signature map and with

$$\Phi_M : \mathcal{P}^p \to \mathbf{T}_1(V), \quad \Phi_M = \Lambda \circ \mathrm{S}_{\leq M}$$

the normalized truncated signature. Similarly, we define

$$\mathrm{k}_M(x, y) := \langle \Phi_M(x), \Phi_M(y) \rangle \ .$$

In complete analogy, we define Φ_M^φ, resp. k_M^κ, as the lifts of state space features φ, resp. state space kernel κ, to pathspace, see Section 5.5 and Section 7.4.

Even if V is finite dimensional, the truncated normalized signature Φ_M has up to $\dim(V)^M$ non-zero coordinates which makes a direct computation for already moderately high-dimensional V infeasible. On the other hand, below we show that the kernel k_M can be efficiently computed, by adapting the recurivse algorithms in [33]. The complexity only relies on the cost of evaluating the state space kernel κ and not the dimension of V.

Secondly, we usually do not have access to a full path sample $x = (x(t))_{t\in[0,1]}$, but only a finite samples of its values, $x^\pi = (x(t_i))_i$ for t_i in a partition $\pi = \{0 \leq t_1 < \cdots \leq 1\}$ of $[0, 1]$. This partition π might even change from path to path. Hence, in addition to truncation at a fixed degree, we need a kernel $\mathrm{k}_M^{\kappa,+}$ on the domain of sequences in $\mathcal{X}$, such the difference

$$\left| \mathrm{k}_M^\kappa(x, y) - \mathrm{k}_M^{\kappa,+}(x^\pi, y^{\pi'}) \right|$$

gets small as the mesh of partitions π and π', $\operatorname{mesh} \pi := \max_{t_i\in\pi}(t_{i+1} - t_i)$, goes to 0.

We state the following result only for $p = 1$, but in Appendix G we present very general approximations to rough paths for any $p \geq 1$ with explicit convergence rates. Recall the notation from Definition 5.12.

Proposition 7.10. *Let $M \geq 1$, (H, κ) a RKHS on $\mathcal{X}$. Let $\Lambda : \mathbf{T}_1(H) \to \mathbf{T}_1(H)$ be a tensor normalization arising from a function ψ as in Corollary A.3. Let k_M^κ be the corresponding normalized kernel.*

(1) There exists a bounded, positive definite kernel on the space of sequences in $\mathcal{X}$ of arbitrary length, $\mathcal{X}^+ := \bigcup_{l\geq 0} \mathcal{X}^l$,

$$\mathrm{k}_M^{\kappa,+} : \mathcal{X}^+ \times \mathcal{X}^+ \to \mathbb{R} \ ,$$

such that for $x, y \in C^1(\mathcal{X})$ and partitions π, π' of $[0, 1]$

$$\left| \mathrm{k}_M^\kappa(x, y) - \mathrm{k}_M^{\kappa,+}(x^\pi, y^{\pi'}) \right| \leq O(\Delta_x^\pi + \Delta_y^{\pi'} + \sqrt{\Delta_x^\pi + \Delta_y^{\pi'}}) \ ,$$

where the proportionality constant does not depend on x, y and

$$\Delta_x^\pi := \|x\|_{1\text{-var}} e^{\|x\|_{1\text{-var}}} \max_{t_i\in\pi} \|x\|_{1\text{-var};[t_i,t_{i+1}]} \ .$$

(2) The kernel $\mathrm{k}_M^{\kappa,+}(x^\pi, y^{\pi'})$ can be evaluated in $O(|\pi||\pi'|(c + M) + q)$ time and $O(|\pi||\pi'| + r)$ memory, where $|\pi|$ is the number of time points in π, c is the cost of one evaluation of the kernel κ, and q and r are the total time and memory costs respectively of a single evaluation of ψ and of finding the unique non-negative root of a polynomial $P(\lambda) = \sum_{m=0}^M a_m \lambda^{2m}$ with $a_0 \leq 0 \leq a_1, \ldots, a_M$.

Proof. Point (1) follows from Proposition G.6 by choosing $N = 1$ therein. Point (2) follows from [33, Alg. 3] and the remark that running the algorithm allows one to compute the sequence $(\|(\mathrm{S}_{M,1}^{\mathrm{linear},\pi})^m(x)\|_{H^{\otimes m}})_{m=0}^M$ at no additional cost (where we used notation from Definition G.4) from which the conclusion follows in the same way as the proof of Proposition 4.2. □

Remark 7.11. *Using low-rank approximations as in* [33, Sec. 6.3.2] *the computational cost can be reduced to $O((|\pi| + |\pi'|)(c + M) + q)$ time and $O(|\pi| + |\pi'| + r)$ memory.*

8. Application: two-sample tests for stochastic processes

We apply our results to the two-sample testing problem

$$H_0 : P_X = P_Y \text{ against the alternative } H_1 : P_X \neq P_Y.$$

Here P_X,P_Y denote the laws of stochastic processes $X = (X_t)$, $Y = (Y_t)$.

8.1. **Test statistics.** Denote with $X = (X_t)_{t\in[0,1]}, Y = (Y)_{t\in[0,1]}$ two independent stochastic process carried on a probability space $(\Omega, \Sigma, \mathbb{P})$. We are given m resp. n i.i.d. samples from X resp. Y, which we denote with $X_1, \dots, X_m$ resp. $Y_1, \dots, Y_n$. For brevity, we discuss $m = n$, but the same argument holds for $m \neq n$ with minor modifications.

A test is a measurable subset R such that the null hypothesis H_0 is refuted if $(X_1, \dots, X_m, Y_1, \dots, Y_n) \in R$. Given laws (P_X, P_Y), the probability of falsely rejecting the null is called the type I error, and similarly the probability of falsely accepting the null is called the type II error. If the type I error can be bounded from above, uniformly over all $\mathbb{P}$ under which X and Y are independent, by a constant α, then we say that the test R is of level α. In the MMD context, a natural candidate for a test is

$$R = T^{-1}([c_\alpha, \infty))$$

where the test statistic T approximates $d_k(P_X, P_Y)$ and the threshold c_α is chosen appropriately. The representation

$$d_k(P_X, P_Y)^2 = \mathbb{E}[k(X, X)] - 2\mathbb{E}[k(X, Y)] + \mathbb{E}[k(Y, Y)]$$

can be discretized in several ways to arrive at test statistics; [27, 26] show that

$$T_U^2(X_1, \dots, X_m, Y_1, \dots, Y_m) := \frac{1}{m(m-1)} \sum_{i,j:i\neq j} k(X_i, X_j) - \frac{2}{mn} \sum_{i,j} k(X_i, Y_j) + \frac{1}{n(n-1)} \sum_{i,j:i\neq j} k(Y_i, Y_j)$$

is an unbiased estimators for $d_k^2(P_X, P_Y)$ that requires $O(m^2)$ computational time and $O(m)$ storage. In [27, 26] it was shown that rejecting the null if $T_U^2 > c_\alpha$ for $c_\alpha := 4\sqrt{-m^{-1}\log\alpha}$ gives a test of level α. There it was also pointed out that the choice of threshold is conservative and can be improved by using data-dependent bounds. A related approach is to simply apply a permutation test, that is, to calculate (or approximate) $T_{perm} := \frac{1}{(2m)!}\sum_\pi T_U^\pi$, where $T_U^\pi := T_U(Z_{\pi(1)}, \dots, Z_{\pi(m)}, Z_{\pi(m+1)}, \dots, Z_{\pi(2m)})$ with $Z_i = X_i$ for $i \leq m$ and $Z_i = Y_{i-m+1}$ for $i \geq m+1$, and where the sum is taken over all $(2m)!$ permutations π of $\{1, \dots, 2m\}$. We then compare T_{perm} to the actual test statistic $T_{obs} := T_U(X_1, \dots, X_m, Y_1, \dots, Y_m)$.

In the experiments below we report both — the empirical distribution of T_U under the null and the alternative as well as the empirical distribution of T_{perm} and the actual observation T_{obs}. We refer to the MMD testing literature for many more details and improvements [47, 27, 26, 45, 31, 16].

8.2. **Data.** An exhaustive empirical study is beyond the scope of this article, but below we study two elementary and natural examples from stochastic processes:

Simple random walk vs. path-dependent random walk: Fix $w \in \mathbb{Z}$, set $X_0 = Y_0 = 0 \in \mathbb{R}$ and define two sequences $(X_0, X_1, \dots, X_t)$

$$X_s = \sum_{i=1}^{s} I_s \text{ and } Y_s = \begin{cases} \sum_{i=1}^{s} J_s \text{ for } s \notin w\mathbb{Z}, \\ \prod_{i=s-w}^{s-1} J_i \end{cases}$$

where $I_1, J_1, I_2, J_2, \dots$ are iid and $\mathbb{P}(I_1 = 1) = \mathbb{P}(I_1 = -1) = 0.5$. That is, X is simple random walk in one dimension, however, for Y every w-th increment is a deterministic function of the $w-1$ preceding increments. See Figure 1 to see samples from X and Y.

Signal vs. noise: A classical problem in signal processing is to test whether a signal perturbed by noise or just noise is observed. A common model is

$$X_t = f_t + \sigma\xi_t\,, \quad \text{and} \quad Y_t = \sigma\xi_t$$

with ξ a d-dimensional noise and $f \in C([0,1], \mathbb{R}^d)$ is a signal. We used a signal

$$f_t = \Big(O_x + r\cos(2\pi kt), O_y + r\sin(2\pi kt)\Big)$$

which spins $k = 10$ times around a circle of radius $r = 0.8$ with random origin $O = (O_x, O_y) \sim N(0, 25I_2)$ and perturbed by an iid Gaussian noise $\xi_t \sim N(0, I_2)$ with volatility $\sigma = 0.5$. See Figure 2 to see samples from X for different values of σ.

We used the same time-discretization, $t_i = \frac{i}{100}$ for $i = 0, \dots, 100$, for all sample paths. This allows us to regard a path sample $(x(t_i))_{i=0,\dots,100}$ as vector in $\mathbb{R}^{101}$ and apply as simple baseline the standard MMD of a kernel on $\mathbb{R}^{101}$ (we used the Gaussian kernel). Choosing hyperparameters is a subtle topic in the kernel two-sample test and we use the usual median heuristic, see [26].

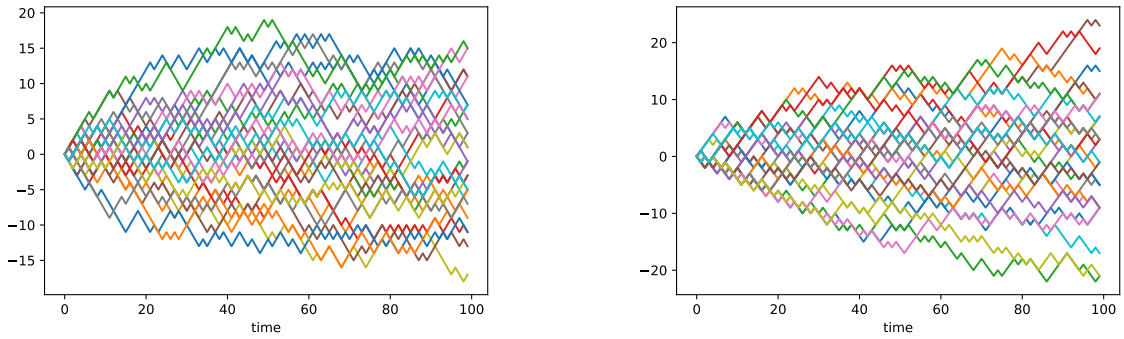


FIGURE 1. The process $X = (X_t)_{t=0,\dots,100}$ is a simple random walk with $X_0 = 0$, $X_t = \sum_{s=0}^{t} I_s$ with I_s i.i.d. Bernoulli $\mathbb{P}(I = -1) = \mathbb{P}(I = 1) = 0.5$. The plot on the left shows 30 independent samples from X. The plot on the right shows 30 independent samples from Y where we used $w = 3$, that is every third increment is the product of the previous two. This elementary example is not trivial if w is unknown: while there are clear patterns in Y, the one-dimensional marginals of the stochastic processes X and Y are the same.

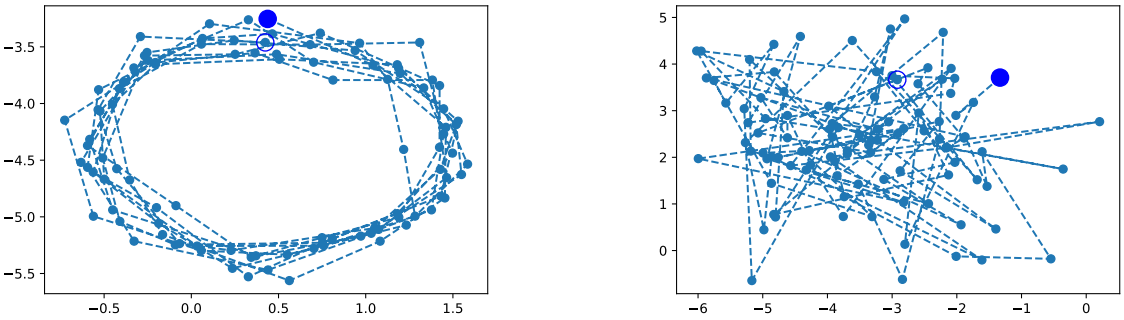


FIGURE 2. The process $X = (X_t)_{t=0,0.1,0.2,\dots,10}$ is given as $X_t = f_t + \sigma N_t$ where the "signal" f is spinning ten times around a circle $f_t = \big(O_x + r\cos(2\pi t), O_y + r\sin(2\pi t)\big)$ with random origin $O = (O_x, O_y) \sim N(0, 25I_2)$ and perturbed by "noise" $N_t \sim N(0, I_2)$. The starting point X_0 is plotted as a filled large circle, the end point X_{10} as a unfilled large circle. The plot on the left shows a sample from X with parameters $r = 1$ and $\sigma = 0.3$, the plot on the right shows a sample from X with $r = 1$ and $\sigma = 1.0$ Note that the process $Y = (Y_t)_{t=0,0.1,\dots,10}$ is given as $Y_t = \sigma N_t$, that is, Gaussian noise. In our experiments we used $\sigma = 0.5$ and $r = 0.8$.

8.3. **Results.** We calculated for our two data sets the test statistics T_U^2 and T_{perm} for different choices of kernels: Gauss, signature-Gauss, signature-Euclidean (for the signature kernels we added lags[9] and for the hyperparameter in the Gauss kernel we used the median heuristic). More precisely, for each of the two testing problems, Figures 1 and 2, we repeated the following 10^3 times: generate $m = 50$ samples from (X, Y) under H_0 (that is X, Y both simple random walks or both signals) and 50 samples under H_1 (that is X a simple random walk resp. signal and Y a path-dependent random walk resp. pure noise). This yields a histograms for T_U^2 under H_0 and one under H_1. A suitable MMD for the problem should produce a large difference in the support of these empirical distributions. The results are shown as the first two plots in each row of Figures 3 and 4. Additionally, we generate a histogram of T_{perm} (approximated with 250 permutations) by generating 50 samples from (X, Y) under H_0, and plot the value of the actual test statistic T_{obs} (red); this is the rightmost plot in each row. Finally, we also tested the robustness of our approach to deal with paths samples of different length; the results are shown in Figure 5. Notable findings are

Gauss kernel: Neither the empirical distribution of the T_U^2 estimator or the permutation test T_{perm} picks up any difference for the random walk experiment. This is not suprising, since testing power generically decreases

[9]By adding l lags, we mean the map $(x_1, \dots, x_n) \mapsto (\bar{x}_1, \dots, \bar{x}_n)$ where $\bar{x}_i = (x_i, x_{i+1}, \dots, x_{i+l}) \in \mathbb{R}^{d+dl}$ for $i \le n - l$ and $\bar{x}_n = (x_n, x_n, \dots, x_n)$, $\bar{x}_{n-1} = (x_{n-1}, x_n, \dots, x_n)$, $\bar{x}_{n-2} = (x_{n-2}, x_{n-1}, x_n, \dots, x_n)$, and so forth. Adding lags is classic in time series analysis, but it comes at the cost of working in higher dimensional state spaces, i.e. $\mathbb{R}^{d+dl}$ instead of $\mathbb{R}^d$.

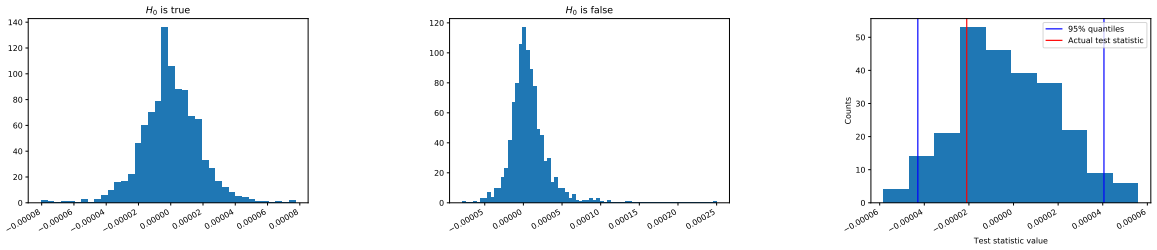
in dimension [43] and the statistical difference between the processes is in the order structure of increments. It does slightly better for the the signal versus noise experiment where it shows a different location for T_U^2, but the supports have a big overlaps, nevertheless the permutation test places the actual test statistic slightly outside the 95% quantile.

Euclidean signature kernel: The Euclidean signature kernels picks up a little difference between the empirical distribution of T_U^2 for the random walk experiment where the tails under H_1 get a bit larger. However, they still have a big overlap. Similarly, the permutation test does not pick up a difference. On the other, for the signal vs noise example, this kernel gives very strong results: the support of T_U^2 under null and alternative are disjoint.

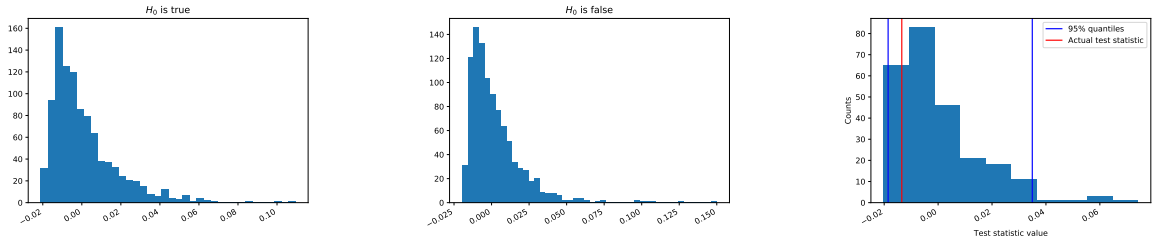
Gauss signature kernel: This is the only kernel that almost completely separated the support of the test statistic T_U^2 under the null and the alternative for the random walk data; furthermore, the permutation test places the actual test statistic far outside the 95% quantile. Similarly, it gives good results for the signal vs noise experiment. There it performs slightly worse than the Euclidean signature kernel.

Data dependent threshold: For all three kernels, the simple two-sample test that rejects if $T_U^2 > c_\alpha$ does not reject the null for a confidence level of 95% in nearly all of the 10^3 repetitions. Although some kernels show a clear difference in the histograms, to reject the null with the threshold $c_\alpha = O(m^{-0.5})$, many more than $m = 50$ samples would be needed. Already in [26] it was noted that the naive threshold might often be too conservative and subsequently more sophisticated approaches were developed to optimize the power of MMD tests which we did not apply here.

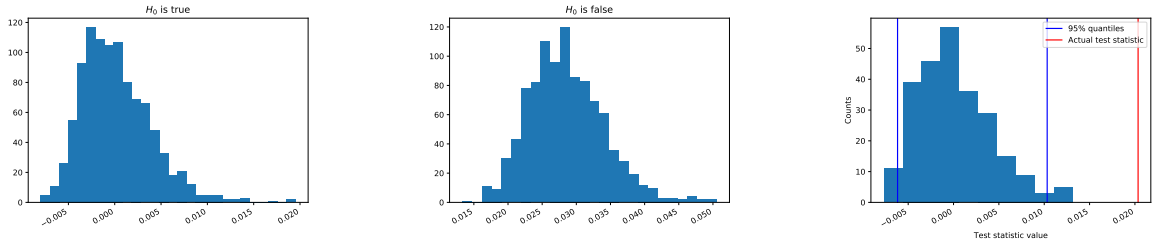
Robustness in tick length: When the number of ticks in each observation varies, it is difficult to use kernels for vector-valued data since the vectors are of different dimensions. On the other hand, signature kernels make such sequences of different length comparable. As Figure 5 shows, despite losing much of the tick data, there is enough structure such that the Gauss-signature kernel picks up essentially the same statistical significance as the corresponding result in Figure 3 that uses no downsampling: the support of T_U^2 is almost disjoint under null and alternative and the permutation test places the actual test statistic again outside the 95% quantile.



(A) The Gauss kernel $\mathrm{k}(x, y) = \exp(-\gamma\|x - y\|^2)$ with the median heuristic for γ. Here, $x \in \mathbb{R}^{101}$ denotes the vector given by ordering the path observations over time as vector $(X_0, X_1, \ldots, X_{100})$.

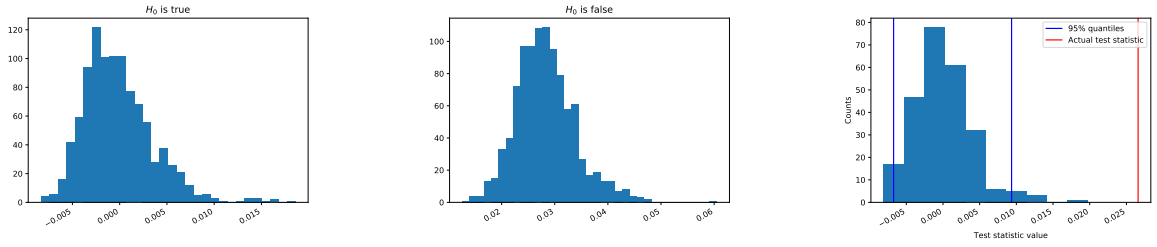


(B) The signature kernel $\mathrm{k}^+_{\kappa,M}(x, y)$ where $M = 4$, κ is the Euclidean kernel, and 4 lags are used.



(C) The signature kernel $\mathrm{k}^+_{\kappa,M}(x, y)$ where $M = 4$, κ is the Gauss kernel, and 4 lags are used.

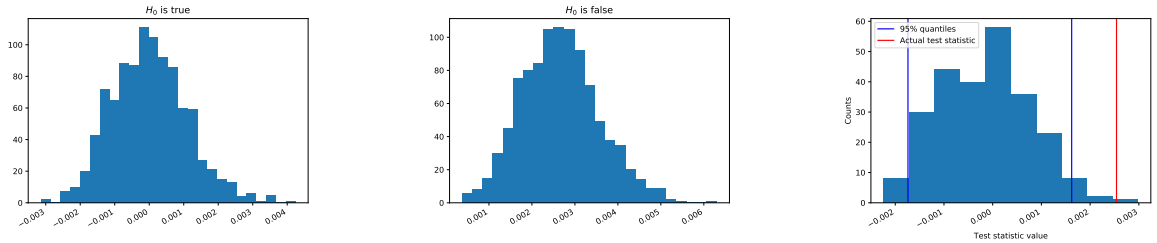
FIGURE 3. **Simple random walk vs path-dependent random walk.** We used the random walk vs path-dependent random walk data, Figure 1. The first two plots on the left in each row show the empirical density of the unbiased estimator T_U^2 under H_0 and H_1. The plots on the right show the histogram of T_{perm} (approximated using 250 permutations) and the value of the actual test statistic T_U (red).



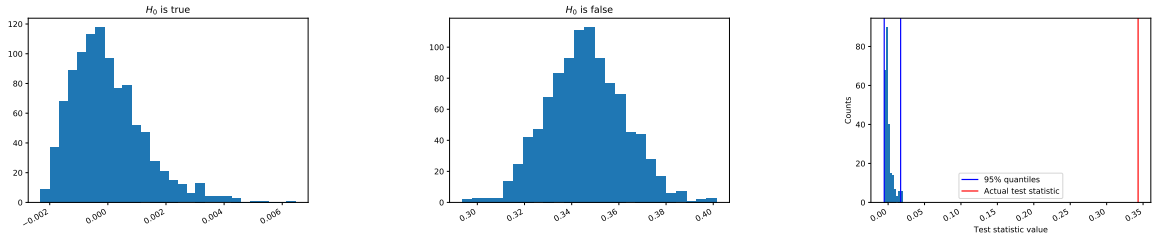
(A) The signature kernel $\mathrm{k}^+_{\kappa,M}(x, y)$ where $M = 4$, κ is the Gauss kernel, and 4 lags are used.

FIGURE 5. **Robustness of path samples of different length** We used the random walk vs path-dependent random walk data, Figure 1. We downsampled each of the paths from 101 ticks to $80 - 101$ ticks by deleting ticks uniformly at random. Despite losing much tick data, the Gauss-signature kernel picks up a difference between null and alternative.

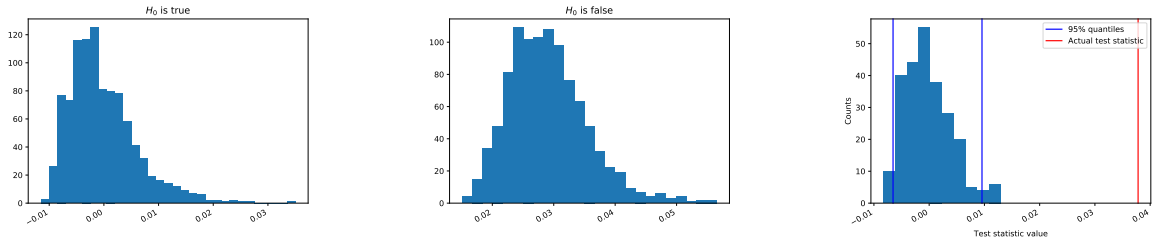
Remark 8.1. *We implement the experiments in Python. For general kernel MMD calculation we use code from the DS3 summer school course "Representing and comparing probabilities with kernels" provided by H. Strathmann*



(A) The Gaussian kernel $\mathrm{k}(x, y) = \exp(-\gamma\|x - y\|^2)$ with the median heuristic for γ.



(B) The signature kernel $\mathrm{k}^+_{\kappa,M}(x, y)$ where $M = 4$, κ is the Euclidean kernel, and 4 lags are used.



(C) The signature kernel $\mathrm{k}^+_{\kappa,M}(x, y)$ where $M = 4$, κ is the Euclidean kernel, and 4 lags are used

FIGURE 4. **Signal vs noise.** We generated 10^4 samples from (X, Y) under H_0 and under H_1 where the signal X and the noise Y are as specified in Figure 2. The histograms on the left [resp. right] show the empirical density of the unbiased estimator MMD under H_0 [resp. under H_1].

.

and D. Sutherland.[10] *For the signature kernels we used code from* [33]. *The signature tensor normalization was performed using a function ψ as in Corollary A.3 of the form*

$$\psi(\sqrt{x}) = \begin{cases} x & \text{if } x \le M\,, \\ M + M^{1+a}(M^{-a} - x^{-a})/a & \text{if } x > M\,, \end{cases}$$

where a, M are constants (chosen as $M = 4$ and $a = 1$ in our experiments). Note that ψ is 1-Lipschitz and is bounded above by $M(1 + \frac{1}{a})$. To find the non-negative root of the polynomial $P(\lambda)$ from part (2) of Proposition 7.10 we used an optimization of Brent's method [6, Ch. 3-4] *implemented as* `optimize.brentq` *in the* `SciPy` *package* [32].

9. SUMMARY AND OUTLOOK

We generalized the classical moment map from the domain of random variables in finite dimensions to the domain of path-valued random variables. This yields a universal and characteristic feature map for stochastic processes, that can be kernelized, and the associated MMD becomes a metric for laws of stochastic processes that can be efficiently estimated from finite samples. Let us highlight that our non-parametric approach of studying probability measures on $C([0, 1], \mathcal{X})$ encompasses a large class of potential examples important to different communities, e.g.

- genuinely discrete sequential data such as text,[11]

[10] available at `https://github.com/karlnapf/ds3_kernel_testing`

[11] If $\mathcal{X}$ is linear, sequences in $\mathcal{X}$ embed into paths in $\mathcal{X}$, $\mathcal{X}^n \hookrightarrow C([0, 1], \mathcal{X})$, via the Donsker embedding $(x_i)_{i=1}^n \mapsto \left(t \mapsto x_{\lfloor nt \rfloor} + (nt - \lfloor nt \rfloor) \sum_{i=1}^{\lceil nt \rceil} x_i\right)$. This pushes measures on sequences to measures on paths and allows to treat genuinely discrete data in our framework; e.g. for text, $\mathcal{X}$ is the free vector space spanned by letters and the path associated with a text (a sequence of letters) is a lattice path; multi-variate time series are sequences in $\mathcal{X} = \mathbb{R}^d$, etc.

- classical time series such as (G)ARCH, ARMA, etc. as used in econometrics,
- cash register and turnstile streams as studied in the data mining/streaming communities,
- continuous semimartingales such as Itô diffusions as studied in mathematical finance or biology,
- stochastic processes in space and time, e.g. stochastic partial differential equations seen as evolution equations (ODEs with noise in infinite-dimensional state space),
- barcodes as they arise in topological data analysis [12],
- the evolution of structured, non-Euclidean objects, such as networks, molecules, images, which are typically studied in machine learning, provided a kernel for the state space $\mathcal{X}$ in which they evolve is given.

Of course, the empirical usefulness of our approach in each of these domains needs to be explored and compared against domain specific state-of-the-art methods; see e.g. [12] where it achieves state-of-the-art performance in standard datasets from topological data analysis. From a more theoretical perspective, the existence of a universal and characteristic feature map for stochastic processes opens several research venues; some of them are

(Kernel) mean embeddings: Once a universal resp. characteristic kernel is given, one can directly apply the well-developed mean embedding literature, see [38] for a recent survey. Among the many possible applications are goodness-of-fit tests or independence testing on pathspace.

Non-independent observations: We assumed that our path-valued samples are independent. This is a reasonable assumption for many real-world applications (e.g. each sample is patient data collected over time). An interesting question is how this assumption can be weakened. For example, often the data consists of a single trajectory and one would like to understand the structure of such an evolution over different time intervals.

Statistics in high/infinite dimensions: Non-parametric testing in high dimensions suffers from decreasing test power [43]. Indeed, the Gauss kernel performed poorly in our experiments, but the normalized signature kernels performed well. This is not too surprising from a stochastic analysis perspective since iterated integrals provide a natural basis for functions of paths (e.g. stochastic Taylor or Wiener chaos type expansions). Finally, we note that signatures are classical in stochastic analysis, but applications in statistics are much more recent, e.g. [41] use it for SDE parameter estimation and it would be interesting to connect this with the methods developed in this paper.

Non-commutative algebra and change of basis: Our strongest results are given by first lifting the path to a high-dimensional space and subsequently computing inner products of signatures using a kernel trick. In a classification context, the same phenomen was emprically observed in [12, 33] where the Gauss-signature kernel generically outperforms the Euclidean-signature kernel. As discussed in Section 5.5, the reason for such an improvement is that more non-linearities are provided. On a more fundamental level, this leads to algebraic questions about "choosing a basis" of the tensor algebra which is better suited for the distribution of the data at hand.

(Semi)-parametric statistics: Often it is justified to restrict attention to a class of stochastic processes, e.g. to (semi)martingales in finance. In such situations, it can be possible to derive bounds on the growth of the iterated integrals such that the unnormalized expected signature characterizes the law of X, see [15]. Even explicit formulas for the expected signature are known (as for Brownian motion in Example 1.1 or more generally Lévy processes [20]) or can be described by a PDE [40]. Our results — a universal and characteristic feature map/kernel giving a computable metric for laws of stochastic processes — then apply even without using the normalization Λ.

Appendix A. Normalization estimates

We follow the notation of Section 3. Let V be a Banach space throughout this section.

Lemma A.1. *Suppose that $\lambda \geq 0$ and $\mathbf{t} \in \mathbf{T}(V)$ such that $\delta_\lambda \mathbf{t} \in \mathbf{T}(V)$. Then $\|\delta_\lambda \mathbf{t} - \mathbf{t}\|^2 \leq |\|\delta_\lambda \mathbf{t}\|^2 - \|\mathbf{t}\|^2|$.*

Proof. We may suppose without loss of generality that $\lambda \geq 1$ (otherwise we swap $\mathbf{t}$ with $\delta_\lambda \mathbf{t}$ and λ with λ^{-1}). Taking derivatives in $\bar{\lambda}$ at some $\bar{\lambda} \in [1, \lambda)$,

$$\frac{d}{d\bar{\lambda}} \|\delta_{\bar{\lambda}} \mathbf{t} - \mathbf{t}\|^2 = \sum_{m=1}^{\infty} 2m(\bar{\lambda}^{2m-1} - \bar{\lambda}^{m-1}) \|\mathbf{t}^m\|_{V^{\otimes m}}^2 \leq \sum_{m=1}^{\infty} 2m\bar{\lambda}^{2m-1} \|\mathbf{t}^m\|_{V^{\otimes m}}^2 = \frac{d}{d\bar{\lambda}} \|\delta_{\bar{\lambda}} \mathbf{t}\|^2 ,$$

where all sums are convergent by the assumption that $\sum_{m=1}^{\infty} \lambda^{2m} \|\mathbf{t}^m\|^2 < \infty$. Since both derivatives are positive and the first is bounded above by the second, the conclusion follows. □

Proposition A.2. *Let $\psi : [1, \infty) \to [1, \infty)$ with $\psi(1) = 1$. For $\mathbf{t} \in \mathbf{T}_1(V)$, let $\lambda(\mathbf{t}) \geq 0$ denote the unique non-negative number such that $\|\delta_{\lambda(\mathbf{t})} \mathbf{t}\|^2 = \psi(\|\mathbf{t}\|)$. Define*

$$\begin{aligned} \Lambda : \mathbf{T}_1(V) &\to \mathbf{T}_1(V) , \\ : \mathbf{t} &\mapsto \delta_{\lambda(\mathbf{t})} \mathbf{t} . \end{aligned}$$

Denote further $\|\psi\|_\infty = \sup_{x \geq 1} \psi(x)$.

(i) It holds that Λ takes values in the set $\{\mathbf{t} \in \mathbf{T}_1(V) : \|\mathbf{t}\| \leq \sqrt{\|\psi\|_\infty}\}$.
(ii) If ψ is injective, then so is Λ.
(iii) Suppose that $\sup_{x \geq 1} \psi(x)/x^2 \leq 1$, $\|\psi\|_\infty < \infty$, and that ψ is K-Lipschitz for some $K > 0$. Then

$$\|\Lambda(\mathbf{s}) - \Lambda(\mathbf{t})\| \leq (1 + K^{1/2} + 2\|\psi\|_\infty^{1/2})(\|\mathbf{s} - \mathbf{t}\|^{1/2} \vee \|\mathbf{s} - \mathbf{t}\|) .$$

Proof. (i) is clear by construction. For (ii) suppose that ψ is injective and that $\Lambda(\mathbf{t}) = \Lambda(\mathbf{s})$. If $\mathbf{t} = \mathbf{1}$, then evidently $\mathbf{s} = \mathbf{t}$. We thus suppose $\mathbf{t} \neq \mathbf{1}$. By definition of Λ, it holds that $\mathbf{s} = \delta_\lambda \mathbf{t}$ for some $\lambda \geq 0$. On the other hand, we also have $\psi(\|\mathbf{t}\|) = \psi(\|\mathbf{s}\|)$. Since ψ is injective and $\lambda \mapsto \|\delta_\lambda \mathbf{t}\|$ is strictly increasing, it follows that $\mathbf{s} = \mathbf{t}$, which proves (ii).

For (iii), suppose $\|\mathbf{s} - \mathbf{t}\| = \epsilon$. Note that $\psi(x) \leq x^2$ implies $\lambda(\mathbf{t}) \leq 1$, and thus $\delta_{\lambda(\mathbf{t})}$ is 1-Lipschitz on $\mathbf{T}_1(V)$. It follows that

$$\begin{aligned} \|\delta_{\lambda(\mathbf{s})} \mathbf{s} - \delta_{\lambda(\mathbf{t})} \mathbf{t}\| &\leq \|\delta_{\lambda(\mathbf{t})} \mathbf{s} - \delta_{\lambda(\mathbf{t})} \mathbf{t}\| + \|\delta_{\lambda(\mathbf{s})} \mathbf{s} - \delta_{\lambda(\mathbf{t})} \mathbf{s}\| \\ &\leq \epsilon + |\|\delta_{\lambda(\mathbf{s})} \mathbf{s}\|^2 - \|\delta_{\lambda(\mathbf{t})} \mathbf{s}\|^2|^{1/2} \\ &\leq \epsilon + |\psi(\|\mathbf{s}\|) - \psi(\|\mathbf{t}\|)|^{1/2} + |\|\delta_{\lambda(\mathbf{t})} \mathbf{t}\|^2 - \|\delta_{\lambda(\mathbf{t})} \mathbf{s}\|^2|^{1/2} \\ &\leq \epsilon + K^{1/2} \epsilon^{1/2} + 2\|\psi\|_\infty^{1/2} \epsilon^{1/2} , \end{aligned}$$

where in the second line we used Lemma A.1, and in the fourth line we used that ψ is K-Lipschitz and without loss of generality that $\lambda(\mathbf{t}) \leq \lambda(\mathbf{s})$ (so that $\|\delta_{\lambda(\mathbf{t})} \mathbf{s}\| \leq \|\delta_{\lambda(\mathbf{s})} \mathbf{s}\| \|\psi\|_\infty^{1/2}$). □

Corollary A.3. *Let $\psi : [1, \infty) \to [1, \infty)$ be injective satisfying $\psi(1) = 1$ and the conditions of item (iii) of Proposition A.2. Then Λ constructed in Proposition A.2 is a tensor normalization.*

Appendix B. Tree like paths

For a topological space $\mathcal{X}$ and a function $x : [0, 1] \to \mathcal{X}$, let $\overleftarrow{x} : [0, 1] \to \mathcal{X}$, $\overleftarrow{x}(t) = x(1 - t)$ denote the time-reversal of x. For another function $y : [0, 1] \to \mathcal{X}$, we denote the concatenation of x with y by

$$x * y : [0, 1] \to \mathcal{X} , \quad x * y(t) = \begin{cases} x(2t) & \text{if } t \in [0, 1/2] \\ y(2t - 1) & \text{if } t \in (1/2, 1] . \end{cases}$$

Note that $x * y$ is a continuous path if x and y are continuous and $x(1) = y(0)$. A function $x : [0, 1] \to \mathcal{X}$ is called tree-like if x is continuous and there exists an $\mathbb{R}$-tree $\mathfrak{T}$, a continuous function $\phi : [0, 1] \to \mathfrak{T}$, and a map $\psi : \mathfrak{T} \to \mathcal{X}$ such that $\phi(0) = \phi(1)$ and $x = \psi \circ \phi$. We say that x and y are tree-like equivalent if $x * \overleftarrow{y}$ is tree-like.

Appendix C. Feature for geometric rough paths

To define the iterated integrals $\int dx^{\otimes m}$ using Riemann–Stieltjes–Young integration requires $x : [0,1] \to V$ to have finite p-variation for some $p \in [1,2)$. In particular, many processes of interest in stochastic analysis fall outside this scope. The rough paths approach is to flesh out the abstract properties of maps that associate with a path x over a time interval $[s,t]$, an element of $\prod_{m=0}^{\lfloor p \rfloor} V^{\otimes m}$ such that it "behaves like" $\int dx^{\otimes m}$. For a large class of examples such as semimartingales, Gaussian processes, SPDEs, etc., it is possible to find such maps and ultimately associate with almost every sample path of a process an element $(\int dx^{\otimes m})_{m\geq 0} \in \prod_{m\geq 0} V^{\otimes m}$.

Throughout this appendix, let V be a Banach space. Observe that $\prod_{m=0}^{M} V^{\otimes m}$ is an algebra by extending tensor multiplication linearly

$$\left(\mathbf{s}^0, \mathbf{s}^1, \dots\right) \otimes (\mathbf{t}^0, \mathbf{t}^1, \dots) = \left(\mathbf{s}^0 \otimes \mathbf{t}^0, \mathbf{s}^0 \otimes \mathbf{t}^1 + \mathbf{s}^1 \otimes \mathbf{t}^0, \mathbf{s}^0 \otimes \mathbf{t}^2 + \mathbf{s}^1 \otimes \mathbf{t}^1 + \mathbf{s}^2 \otimes \mathbf{t}^0, \dots\right),$$

i.e. for $m \leq M$, the $V^{\otimes m}$-coordinate equals $\sum_{n=0}^{m} \mathbf{s}^{m-n} \otimes \mathbf{t}^n$. If $x \in C^1$, then the classically defined integrals $\int dx^{\otimes m}$ satisfy the so-called Chen's identity

$$\left(1, \int_s^t dx, \int_s^t dx^{\otimes 2}, \dots\right) \otimes \left(1, \int_t^u dx, \int_t^u dx^{\otimes 2}, \dots\right) = \left(1, \int_s^u dx, \int_s^u dx^{\otimes 2}, \dots\right), \tag{12}$$

or simply

$$\mathbf{x}(s,t) \otimes \mathbf{x}(t,u) = \mathbf{x}(s,u)\ ,$$

where we denote $\mathbf{x}(s,t) := (1, \mathbf{x}^1(s,t), \mathbf{x}^2(s,t), \dots) := (1, \int_s^t dx, \int_s^t dx^{\otimes 2}, \dots) \in \prod_{m=0}^{M} V^{\otimes m}$.

Definition C.1. *Let $p \geq 1$. A p-rough path is a continuous map*

$$\mathbf{x} = (\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^{\lfloor p \rfloor}) : [0,1]^2 \to \prod_{m=0}^{\lfloor p \rfloor} V^{\otimes m}$$

such that

(a) $\mathbf{x}$ is multiplicative i.e. $\mathbf{x}^0 \equiv 1$, $\mathbf{x}(t,t) = (1,0,\dots,0)$, and

$$\mathbf{x}(s,t) \otimes \mathbf{x}(t,u) = \mathbf{x}(s,u), \quad \forall (s,t),(t,u) \in [0,1]^2, \tag{13}$$

(b) for all $0 \leq m \leq \lfloor p \rfloor$

$$\sup_{\substack{n\geq 1 \\ 0\leq t_1 \leq \dots \leq t_n \leq 1}} \sum_{i=1}^{n-1} \|\mathbf{x}^m(t_i, t_{i+1})\|^{p/m} < \infty. \tag{14}$$

A p-rough path $\mathbf{x}$ is called geometric if there exists a sequence of smooth rough paths $(\mathbf{x}_n)_{n\geq 1}$ such that $\mathbf{x}_n \to \mathbf{x}$ in p-variation metric.[12] *We denote by $C^p = C^p(V)$ the space of geometric p-rough paths equipped with p-variation metric.*

If $x \in C([0,1], V)$ and there exists a p-rough path $\mathbf{x}$ such that

$$\mathbf{x}^1(s,t) = x(t) - x(s), \quad \forall (s,t) \in [0,1]^2,$$

then we call $\mathbf{x}$ a p-rough path lift of x.

Property (a) is the abstract version of Chen's identity (12). Property (b) is natural, since for a bounded variation path x, i.e. $p = 1$, we have

$$\sup_{\substack{n\geq 1 \\ 0\leq t_1 \leq \dots \leq t_n \leq 1}} \sum_{i=1}^{n-1} \|\mathbf{x}^m(t_i, t_{i+1})\|^{1/m} < \infty \quad \text{where} \quad \mathbf{x}^m(s,t) := \int_s^t dx^{\otimes m},$$

i.e. integration adds one more degree of regularity.

Note that every $\mathbf{x} \in C^p$ gives rise to a unique path (which we denote with the same symbol) $\mathbf{x} : t \mapsto \mathbf{x}(0,t)$ with initial point $\mathbf{x}(0) = \mathbf{1}$. Conversely, property (a) allows us to recover $\mathbf{x}(s,t)$ as $\mathbf{x}(s,t) = \mathbf{x}(s)^{-1} \otimes \mathbf{x}(t)$. Hence we can equivalently treat $\mathbf{x}$ as a $\prod_{m=0}^{\lfloor p \rfloor} V^{\otimes m}$-valued path.

Remark C.2.

- *In the case $V = \mathbb{R}^d$, a p-rough path satisfies the shuffle identity* (11) *if it is geometric (and the converse is almost true, see* [23, Sec. 9.2]*).*
- *While we work here only with continuous paths, which is the usual setting for rough paths theory, we note that meaningful extensions exist to the discontinuous setting* [14, 11, 20, 49, 24].

[12] See [36, Eq. (3.70)] for the definition of the p-variation metric; for brevity, we do not give a definition here, particularly because its explicit form plays no role in the sequel.

Many stochastic processes of interest admit lifts to geometric p-rough paths (for a.e. sample trajectory).

Example C.3.

(1) ***Bounded variation paths.*** *For $p = 1$, we recover the setting of Section 5 since every $x \in C^1$ is a 1-rough path by identifying $\prod_{m=0}^{1} V^{\otimes m}$ with V.*
(2) ***(Semi)martingales.*** *With $2 < p < 3$ we cover continuous (semi)martingale theory: every continuous semimartingale $X : [0,1] \to \mathbb{R}^d$ is of bounded p-variation for any $p > 2$. Stochastic (Itô or Stratonovich) integration can be used to the define the first $2 = \lfloor p \rfloor$ iterated integrals*

$$\mathbf{X}(s,t) = \Big(1, \int_s^t dX, \int_s^t dX^{\otimes 2}\Big).$$

One can verify that (13) *and* (14) *hold almost surely, hence $\mathbf{X}$ is a p-rough path. Moreover, if Stratonovich integration is used, then $\mathbf{X}$ is geometric.*
(3) ***Gaussian processes.*** *Many Gaussian processes admit canonical lifts to geometric p-rough paths. A simple criterion for such a lift to exist can be given in terms of the covariance function* [17, 23, 22]*, which in particular covers fractional Brownian motion with Hurst parameter $H > 1/4$.*
(4) ***Markov processes.*** *Likewise, many Markov processes admit canonical lifts to geometric p-rough paths (typically $2 < p < 3$). These include diffusions on fractals and Markov processes arising from uniformly elliptic Dirichlet forms* [1, 34, 21, 13]*. Such processes are typically not semimartingales and thus fall outside the scope of Itô–Stratonovich calculus.*

Remark C.4. *Branched rough paths* [28] *are a generalization of geometric rough paths for which the chain rule is not required to hold. Such considerations arise naturally when dealing with Itô-type calculus* [29] *or certain renormalization procedures* [7]*. We indicate in Appendix D how our considerations extend to this setting.*

C.1. **Ordered moments for rough paths.** A key result of Lyons [35] is that the signature map S defined by (8) extends to the space of p-rough paths: for any $\mathbf{x} \in C^p$, the integrals $\int d\mathbf{x}^{\otimes m}$ are canonically defined for all $m > \lfloor p \rfloor$. Moreover, S is injective on C^p up to tree-like equivalence [4]. We summarise these results in the following extension of Theorem 5.3.

Theorem C.5. *Let $p \geq 1$. There exists a map $\mathrm{S} : C^p \to \mathbf{T}_1(V)$, $\mathrm{S}(\mathbf{x}) = (1, \mathrm{S}^1(\mathbf{x}), \mathrm{S}^2(\mathbf{x}), \ldots)$ such that $\mathrm{S}^m(\mathbf{x}) = \mathbf{x}^m(0,1)$ for all $m = 0, \ldots, \lfloor p \rfloor$. Furthermore, if $\mathbf{x}, \mathbf{y} \in C^p$, then $\mathrm{S}(\mathbf{x}) = \mathrm{S}(\mathbf{y})$ if and only if $\mathbf{x} * \overleftarrow{\mathbf{y}}$ is tree-like.*

Example C.6.

- *Let $x \in C^1$. Then $\mathrm{S}(x) \in \mathbf{T}(V)$ defined by* (8) *agrees with $\mathrm{S}(x)$ in Theorem C.5.*
- *Let $X = (X(t))_{t\in[0,1]}$ be a continuous semimartingale in $\mathbb{R}^d$ with geometric p-rough path lift $\mathbf{X} = \big(1, \mathbf{X}^1, \mathbf{X}^2\big) \equiv \big(1, \int dX, \int dX^{\otimes 2}\big)$ as in part (2) of Example C.3. Then*

$$\mathrm{S}(\mathbf{X}) = \Big(1, \int_0^1 dX, \int_0^1 dX^{\otimes 2}, \int_0^1 dX^{\otimes 3}, \ldots\Big) \in \mathbf{T}_1(\mathbb{R}^d),$$

where the stochastic integrals are taken in the Stratonovich sense.

Theorem C.5 suggests the following generalization of Definition 5.5.

Definition C.7. *For $\mathbf{x}, \mathbf{y} \in C^p$ we write $\mathbf{x} \sim_t \mathbf{y}$ if $\mathbf{x} * \overleftarrow{\mathbf{y}}$ is tree-like. We define the space of unparameterised geometric p-rough paths $\mathcal{P}^p$ as the set of equivalence classes $C^p / \sim_t$ equipped with the quotient topology.*

C.2. **Time parameterization.** As in the bounded variation setting of Section 5.3, we can make the map S variant to the parameterization of time by adding a time component. For any $\mathbf{x} \in C^p(V)$, there is a canonical p-rough path lift of $t \mapsto (\mathbf{x}^1(t), t)$, denoted by $\overline{\mathbf{x}} \in C^p(V \oplus \mathbb{R})$, and which extends $\mathbf{x}$ in the sense that $\langle \ell, \mathbf{x}^m \rangle = \langle \ell, \overline{\mathbf{x}}^m \rangle$ for all $m = 0, \ldots, \lfloor p \rfloor$ and $\ell \in (V^{\otimes m})'$, see [36, Sec. 3.3.3]. It clearly holds that $\overline{\mathbf{x}} \sim_t \overline{\mathbf{y}}$ if and only if $\mathbf{x} = \mathbf{y}$.

C.3. **MMD for measures on rough paths.** We conclude this section with the proof of item (4) of Theorem 7.7. Let notation be as in Section 7.3 and Theorem 7.7.

Proposition C.8. *Let $p \geq 1$. Then*

(1) *there exist probability measures μ_n, μ, on $\mathcal{P}^p$ such that $d_{\mathrm{k}}(\mu_n, \mu) \to 0$ but such that μ_n does not converge weakly to μ,*
(2) *if H is finite dimensional, weak convergence of probability measures on $\mathcal{P}^p$ implies convergence in d_{k}.*

The same statements hold with $\mathcal{P}^p$ replaced by C^p.

Proof. (1) Let $p < p' < \lfloor p \rfloor + 1$. Consider $\mathbf{x} \in \mathcal{P}^p$ and a sequence $\mathbf{x}_n \in \mathcal{P}^p$ such that $\mathbf{x}_n$ does not converge to $\mathbf{x}$ as elements of of $\mathcal{P}^p$ but such that $\mathbf{x}_n \to \mathbf{x}$ as elements of $\mathcal{P}^{p'}$, i.e., in the p'-variation metric (it is a simple exercise to construct such a sequence for any $\mathbf{x} \in \mathcal{P}^p$). Due to continuity of $\Phi : \mathcal{P}^{p'} \to \mathbf{T}(H)$, we conclude that $\|\Phi(\mathbf{x}_n) - \Phi(\mathbf{x})\| \to 0$. In particular, for the corresponding Dirac delta measures, we have

$$d_{\mathrm{k}}(\delta_{\mathbf{x}_n}, \delta_{\mathbf{x}}) = \sup_{\mathbf{t} \in \mathbf{T}(H), \|\mathbf{t}\| \le 1} |\langle \mathbf{t}, \Phi(\mathbf{x}_n)\rangle - |\langle \mathbf{t}, \Phi(\mathbf{x})\rangle| \to 0.$$

However, due to the assumption that $\mathbf{x}_n$ does not converge to $\mathbf{x}$ in $\mathcal{P}^p$, it holds that $\delta_{\mathbf{x}_n}$ does not converge weakly to $\delta_{\mathbf{x}}$ as probability measures on $\mathcal{P}^p$.

(2) Let μ and μ_n be probability measures on $\mathcal{P}^p$ such that $\mu_n \to \mu$ weakly. Observe that

$$\begin{aligned} d_{\mathrm{k}}(\mu_n, \mu)^2 &= \sup_{\mathbf{t} \in \mathbf{T}(H), \|\mathbf{t}\| \le 1} \Big| \int_{\mathcal{P}^p} \langle \mathbf{t}, \Phi(\mathbf{x})\rangle \mu_n(d\mathbf{x}) - \int_{\mathcal{P}^p} \langle \mathbf{t}, \Phi(\mathbf{x})\rangle \mu(d\mathbf{x}) \Big|^2 \\ &= \Big\| \int_{\mathcal{P}^p} \Phi(\mathbf{x}) \mu_n(d\mathbf{x}) - \int_{\mathcal{P}^p} \Phi(\mathbf{x}) \mu(d\mathbf{x}) \Big\|^2_{\mathbf{T}(H)} \\ &= \sum_{m \ge 0} \Big\| \int_{\mathcal{P}^p} \lambda(\mathrm{S}(\mathbf{x}))^m \, \mathrm{S}^m(\mathbf{x}) \mu_n(\mathbf{x}) - \int_{\mathcal{P}^p} \lambda(\mathrm{S}(\mathbf{x}))^m \, \mathrm{S}^m(\mathbf{x}) \mu(\mathbf{x}) \Big\|^2_{H^{\otimes m}}. \end{aligned}$$

Note that for every $m \ge 0$, $\mathbf{x} \mapsto \lambda(\mathrm{S}(\mathbf{x}))^m \, \mathrm{S}^m(\mathbf{x})$ is a continuous bounded function from $\mathcal{P}^p$ into the finite dimensional vector space $H^{\otimes m}$. It follows from the weak convergence $\mu_n \to \mu$ that the final quantity converges to zero as $n \to \infty$, which proves (2).

The corresponding claims for C^p follow in an identical manner. □

Appendix D. Branched rough paths

The purpose of this appendix is to demonstrate how the abstract results of Section 6 carry over to the setting of branched rough paths [28]. Branched rough paths generalise geometric rough paths in that they allow a violation of the usual chain rule. In particular, Brownian motion with Itô calculus falls naturally into this framework.

We present the definition and properties of finite-dimensional rough paths; see [29, 7] for further details and [9] for a treatment of infinite-dimensional branched rough paths. Let $\mathcal{B}$ denote the set of rooted, combinatorial trees with node labels from the set $\{1, \dots, d\}$, and $\mathcal{F}$ denote the set of all unordered finite collections of elements of $\mathcal{B}$. We call elements of $\mathcal{F}$ forests. We denote by $\mathbf{1}$ the empty forest in $\mathcal{F}$ (note that $\mathbf{1} \notin \mathcal{B}$). Every element in $\tau \in \mathcal{B}$ can be written in a unique way as $\tau = [\tau_1 \dots \tau_n]_i$, where $\tau_1 \dots \tau_n \in \mathcal{F}$ and $i \in \{1, \dots, d\}$ denotes the label of the root of τ. Note that the possibility $n = 0$ is allowed, in which case $\tau_1 \dots \tau_n = \mathbf{1}$ and $\tau = [\mathbf{1}]_i$ is simply the single labelled node $\bullet_i$.

Definition D.1. *For $\tau \in \mathcal{F}$, let $|\tau|$ denote the number of nodes in τ. For $p \ge 1$, let $\mathcal{F}^{\lfloor p \rfloor}$ denote the set of all $\tau \in \mathcal{F}$ for which $|\tau| \le p$. Let $\mathcal{H}^{\lfloor p \rfloor}$ denote the linear span over $\mathbb{R}$ of $\mathcal{F}^{\lfloor p \rfloor}$ equipped with the inner product $\langle \cdot, \cdot \rangle$ for which $\mathcal{F}^{\lfloor p \rfloor}$ forms an orthonormal basis.*

A branched p-rough path is a map $\mathbf{x} : [0,1]^2 \to \mathcal{H}^{\lfloor p \rfloor}$, which satisfies the algebraic conditions for all $s, t, u \in [0,1]$

(1) $\mathbf{x}(t,t) = \mathbf{1}$ and $\langle \mathbf{x}(s,t), \mathbf{1} \rangle = 1$,
(2) $\langle \mathbf{x}(s,t), \tau_1 \dots \tau_n \rangle = \langle \mathbf{x}(s,t), \tau_1 \rangle \dots \langle \mathbf{x}(s,t), \tau_n \rangle$,
(3) $\mathbf{x}(s,u) = \mathbf{x}(s,t) \star \mathbf{x}(t,u)$, where $\star$ is the Grossman–Larson product,

and the analytic condition

$$\forall \tau \in \mathcal{F}^{\lfloor p \rfloor}, \quad \sup_{\substack{0 \le n \\ 0 \le t_1 \le \dots \le t_n \le 1}} \sum_{i=1}^{n-1} |\langle \mathbf{x}(t_i, t_{i+1}), \tau \rangle|^{p/|\tau|} < \infty. \tag{15}$$

Let $\mathcal{Y}_p$ denote the space of branched p-rough paths.

Example D.2. *Every path $x \in C^1(\mathbb{R}^d)$ can be viewed canonically as a branched p-rough path $\mathbf{x}$ for any $p \ge 1$ as follows. We define $\langle \mathbf{x}(s,t), \tau \rangle$ recursively for $\tau = [\tau_1 \dots \tau_n]_i \in \mathcal{B}$ by*

$$\langle \mathbf{x}(s,t), [\tau_1 \dots \tau_n]_i \rangle = \int_s^t \langle \mathbf{x}(s,u), \tau_1 \dots \tau_n \rangle dx^i(u) \,, \quad \forall s, t \in [0,1] \,, \tag{16}$$

with the base case $\langle \mathbf{x}(s,t), \mathbf{1} \rangle = 1$, and enforcing property (2). This procedure defines $\langle \mathbf{x}(s,t), \tau \rangle$ uniquely for every $\tau \in \mathcal{F}$. For example,

$$\langle \mathbf{x}(s,t), \bullet_i \rangle = \int_s^t 1 dx^i(u) = x^i(t) - x^i(s).$$

More generally, a "ladder" tree of the form

$$\tau = [\dots[[\bullet_{i_1}]_{i_2}]\dots]_{i_m} \tag{17}$$

encodes part of the m-th ordered moment (iterated integral) of x in the interval $[s,t]$

$$\langle \mathbf{x}(s,t), \tau\rangle = \int_{s\le t_1\le\dots\le t_m\le t} dx^{i_1}(t_1)\dots dx^{i_m}(t_m)\ .$$

The multiplicative property (3) follows from Fubini's theorem, and the bound (15) *for every* $p \ge 1$ *is elementary.*

Remark D.3. *Definition D.1 generalizes the space of geometric rough path* $C^p(\mathbb{R}^d)$*. Indeed, there is a canonical algebra embedding of* $(\mathbf{T}(\mathbb{R}^d), \otimes)$ *into the Grossman–Larson algebra of forests* $(\mathcal{H}, \star)$ *under which every geometric rough path is mapped to a branched rough path. See* [7] *for details.*

Example D.4. *Let* $X : [0,1] \to \mathbb{R}^d$ *be a semi-martingale and* $p > 2$*. For* $\tau \in \mathcal{F}^{\lfloor p\rfloor}$ *and* $s \in [0,1]$*, define the* $\mathbb{R}$*-valued semi-martingale* $\langle \mathbf{X}(s,\cdot),\tau\rangle : [s,1] \to \mathbb{R}$ *by properties (1) and (2), and the identity* (16) *where the integral is taken in the sense of Itô. Then* $\mathbf{X} : [0,1]^2 \to \mathcal{H}^{\lfloor p\rfloor}$ *is a.s. a branched p-rough path which in general is non-geometric.*

D.1. **Ordered moments for branched rough paths.** An important property of branched rough paths is an analogue of Theorem C.5. Let $\mathcal{H}^*$ denote the vector space of formal series in forests.

Theorem D.5. *Let* $p \ge 1$*. There exists a map* $\mathrm{S} : \mathcal{Y}_p \to \mathcal{H}^*$*, such that* $\langle \mathrm{S}(\mathbf{x}),\tau\rangle = \langle \mathbf{x}(0,1),\tau\rangle$ *for all* $\tau \in \mathcal{F}^{\lfloor p\rfloor}$*. Furthermore, if* $\mathbf{x},\mathbf{y} \in \mathcal{Y}_p$*, then* $\mathrm{S}(\mathbf{x}) = \mathrm{S}(\mathbf{y})$ *if and only if* $\mathbf{x} * \overleftarrow{\mathbf{y}}$ *is tree-like.*

As before, the map S is called the signature map and its construction follows from the sewing lemma [28]. See [3, Thm. 5.1] for the second statement of Theorem D.5.

As in Section 6, it is possible to construct a universal feature for unparameterized branched rough paths. To avoid introducing normalizations and addition of time-components, we state the following version of Corollary 6.2. See [3, Prop. 5.16] for exact details.

Corollary D.6. *Let* μ_1, μ_2 *be finite Borel measures on* $\mathcal{Y}_p$*. Suppose further that* $\mu_1[\langle \mathrm{S}(\cdot),\tau\rangle]$ *decays sufficiently fast as* $|\tau| \to \infty$*. Then*

$$\mu_1 = \mu_2 \text{ up to tree-like equivalence} \quad \text{iff} \quad \mu_1[\langle \mathrm{S}(\cdot),\tau\rangle] = \mu_2[\langle \mathrm{S}(\cdot),\tau\rangle] \quad \forall \tau \in \mathcal{F}\ .$$

D.2. **Efficient computation.** In this subsection, we demonstrate an inner product on branched rough path signatures which can be recursively computed using a Horner-type scheme. This result can be seen as a branched rough path analogue of [33, Thm. 2]. For notational simplicity, we consider the one-dimensional case $\mathbb{R}^d = \mathbb{R}$ (so we may ignore labels on nodes).

For a forest $\tau \in \mathcal{F}$, define its depth $d(\tau)$ inductively as $d(\mathbf{1}) := 0$, $d(\tau_1\dots\tau_n) = \max\{d(\tau_1),\dots,d(\tau_n)\}$, and

$$d([\tau_1\dots\tau_n]_i) = 1 + \max\{d(\tau_1),\dots,d(\tau_n)\}\ .$$

Consider two paths $x, y \in C^1(\mathbb{R})$ with corresponding branched 1-rough paths $\mathbf{x},\mathbf{y} \in \mathcal{Y}_1$ defined as in Example D.2. For an integer $M \ge 0$, the inner product we consider is

$$\langle \mathrm{S}(\mathbf{x}), \mathrm{S}(\mathbf{y})\rangle_M := \sum_{\substack{\tau\in\mathcal{F}\\ d(\tau)\le M}} c(\tau)\langle \mathrm{S}(\mathbf{x}),\tau\rangle\langle \mathrm{S}(\mathbf{y}),\tau\rangle\ , \tag{18}$$

where $c(\tau)$ is a combinatorial factor to be chosen later.

Proposition D.7. *There exits a choice of constants* $c(\tau)$ *for every* $\tau \in \mathcal{F}$ *such that for every* $M \ge 0$ *and* $x, y \in C^1(\mathbb{R})$*, it holds that*

$$\langle \mathrm{S}(\mathbf{x})_{s,t}, \mathrm{S}(\mathbf{y})_{s,t}\rangle_M = \exp\Big[\int_{(u,v)\in[s,t]^2} \langle \mathrm{S}(\mathbf{x})_{s,u}, \mathrm{S}(\mathbf{y})_{s,v}\rangle_{M-1} dx(s,u)dy(s,v)\Big]\ ,$$

where $\mathrm{S}(\mathbf{x})_{s,t}$ *denotes the image under* S *of the restriction of* $\mathbf{x}$ *to* $[s,t] \subset [0,1]$*.*

Proof. We proceed by induction. The base case $M = 0$ holds trivially with $c(\mathbf{1}) := 1$. For the inductive step, we have

$$\begin{aligned}
&\exp\Big[\int_{(u,v)\in[s,t]^2} \langle \mathrm{S}(\mathbf{x})_{s,u}, \mathrm{S}(\mathbf{y})_{s,v}\rangle_{M-1} dx(s,u)dy(s,v)\Big]\\
&= \exp\Big[\int_{(u,v)\in[s,t]^2} \big[\langle \mathrm{S}(\mathbf{x})_{s,u}, \mathrm{S}(\mathbf{y})(s,v)\rangle_{M-2} + \langle \mathrm{S}(\mathbf{x})_{s,u}, \mathrm{S}(\mathbf{y})(s,v)\rangle_{=M-1}\big] dx(s,u)dy(s,v)\Big]\\
&= \langle \mathrm{S}(\mathbf{x})_{s,t}, \mathrm{S}(\mathbf{y})_{s,t}\rangle_{M-1} \exp\Big[\int_{(u,v)\in[s,t]^2} \langle \mathrm{S}(\mathbf{x})_{s,u}, \mathrm{S}(\mathbf{y})_{s,v}\rangle_{=M-1} dx(s,u)dy(s,v)\Big]\ ,
\end{aligned}$$

where $\langle \mathrm{S}(\mathbf{x})_{s,u}, \mathrm{S}(\mathbf{y})_{s,v}\rangle_{=M-1}$ is defined as in (18) but with the sum restricted to $d(\tau) = M-1$, and where we have used the inductive hypothesis in the final equality. Using the identity $\int_{s\le u\le t}\langle \mathrm{S}(\mathbf{x})_{s,u}, \tau_1\dots\tau_n\rangle dx(s,u) = \langle \mathrm{S}(\mathbf{x})_{s,t}, [\tau_1,\dots,\tau_n]\rangle$, we obtain

$$\exp\Big[\int_{(u,v)\in[s,t]^2}\langle \mathrm{S}(\mathbf{x})_{s,u}, \mathrm{S}(\mathbf{y})_{s,v}\rangle_{=M-1} dx(s,u)dy(s,v)\Big] \quad = \quad \sum_{k=0}^{\infty}\frac{1}{k!}\Big(\sum_{\substack{\tau\in\mathcal{F}\\ d(\tau)=M-1}} c(\tau)\langle \mathrm{S}(\mathbf{x})_{s,t}, [\tau]\rangle\langle \mathrm{S}(\mathbf{y})_{s,t}, [\tau]\rangle\Big)^k \quad .$$

Observe that for every forest $\bar\tau \in \mathcal{F}$ with depth $d(\bar\tau) = M$, the term $\langle \mathrm{S}(\mathbf{x})_{s,t}, \bar\tau\rangle\langle \mathrm{S}(\mathbf{y})_{s,t}, \bar\tau\rangle$ appears once in the following expression (with some combinatorial factor):

$$\langle \mathrm{S}(\mathbf{x})_{s,t}, \mathrm{S}(\mathbf{y})_{s,t}\rangle_{M-1}\sum_{k=1}^{\infty}\frac{1}{k!}\Big(\sum_{\substack{\tau\in\mathcal{F}\\ d(\tau)=M-1}} c(\tau)\langle \mathrm{S}(\mathbf{x})_{s,t}, [\tau]\rangle\langle \mathrm{S}(\mathbf{y})_{s,t}, [\tau]\rangle\Big)^k . \tag{19}$$

It follows that, having chosen $c(\tau)$ for all $d(\tau) \le M-1$, there exists a choice for $c(\bar\tau)$ for each $d(\bar\tau) = M$ such that (19) coincides with $\langle \mathrm{S}(\mathbf{x})_{s,t}, \mathrm{S}(\mathbf{y})_{s,t}\rangle_{=M}$. The conclusion follows by noting that

$$\langle \mathrm{S}(\mathbf{x})_{s,t}, \mathrm{S}(\mathbf{y})_{s,t}\rangle_{M} = \langle \mathrm{S}(\mathbf{x})_{s,t}, \mathrm{S}(\mathbf{y})_{s,t}\rangle_{M-1} + \langle \mathrm{S}(\mathbf{x})_{s,t}, \mathrm{S}(\mathbf{y})_{s,t}\rangle_{=M} .$$

□

Appendix E. Kernel background

Throughout this appendix, we fix a topological space $\mathcal{X}$, an inner product space $(E, \langle\cdot,\cdot\rangle_E)$ and feature map $\Phi : \mathcal{X} \to E$. For $x \in \mathcal{X}$, we define the kernel function $k_x \in \mathbb{R}^{\mathcal{X}}$ by $k_x(y) := k(x,y)$, where $k(x,y) := \langle\Phi(x), \Phi(y)\rangle$. Consider the subspace

$$\mathcal{H}_0 := \mathrm{span}\{k_x : x \in \mathcal{X}\} \subset \mathbb{R}^{\mathcal{X}}$$

equipped with the inner product defined by $\langle k_x, k_y\rangle_{\mathcal{H}_0} := k(x,y)$. We first recall the following theorem of Moore–Aronszajn.

Theorem E.1 (Moore–Aronszajn)**.** *There exits a unique Hilbert space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ with k as the reproducing kernel. Moreover, $\mathcal{H}_0$ is dense in $\mathcal{H}$.*

The following theorem clarifies the relation to our original feature map $x \mapsto \Phi(x)$ and $x \mapsto k_x$ and shows how to construct $\mathcal{H}$ as a subspace of (the completion of) E.

Theorem E.2.

(1) There exists a unique linear map $\Psi : \mathcal{H}_0 \to E$ such that the following diagram commutes:

$$\begin{array}{ccc} \mathcal{X} & \xrightarrow{x\to k_x} & \mathcal{H}_0 \\ & \searrow^{\Phi} & \downarrow{\scriptstyle\Psi} \\ & & E\,. \end{array}$$

(2) The map Ψ is injective and is an isometry onto its image.

(3) Denote by $x \mapsto x'$ the canonical map $E \to E'$ which identifies E with (a subspace of) E'. Then

$$\iota(\Phi(x)') = k_x$$

and the following diagram commutes:

$$\begin{array}{ccccc} \mathcal{X} & \xrightarrow{x\mapsto k_x} & \mathcal{H}_0 & \xhookrightarrow{\mathrm{id}} & \mathbb{R}^{\mathcal{X}} \\ & \searrow^{\Phi} & \downarrow{\scriptstyle\Psi} & & \uparrow{\scriptstyle\iota} \\ & & E & \xrightarrow{x\mapsto x'} & E'\,. \end{array}$$

(4) The image of $\mathcal{H}_0$ in E' under $h \mapsto \Psi(h)'$ is a dense subspace of $\mathrm{Ker}(\iota)^{\perp}$.

Proof. For Point (1), existence and uniqueness of Ψ follows from the observation that if $\sum a_i k_{x_i} \equiv 0$, then $\sum a_i\langle\Phi(x_i), \Phi(y)\rangle = 0$ for all $y \in \mathcal{X}$, and thus $\sum a_i\Phi(x_i)$ is an element of both $\Phi(\mathcal{X})^{\perp}$ and span $[\Phi(\mathcal{X})]$, and thus must be zero. For Point (2), note that $\Psi(\sum a_i k_{x_i}) = 0$ is equivalent to $\sum a_i\Phi(x_i) = 0$, so that

$$\sum a_i\langle\Phi(x_i), \Phi(y)\rangle = \sum a_i k_{x_i}(y) = 0, \quad \forall y \in \mathcal{X}.$$

It follows that Ψ is injective. The fact that Ψ is an isometry follows from Point (1). Point (3) follows immediately since for all $x, y \in \mathcal{X}$

$$\iota(\Phi(x)')(y) = \Phi(x)' \circ \Phi(y) = \langle\Phi(x), \Phi(y)\rangle = k(x,y) = k_x(y).$$

For Point (4), we define for any subset $F \subset E$ the set $F^\circ := \{f' \in E' \mid f'(f) = 0, \forall f \in F\}$. Now $\mathrm{Ker}(\iota)$ consists of all $z \in E'$ such that $z'(\Phi(x)) = 0$ for all $x \in \mathcal{X}$, so that $\mathrm{Ker}(\iota) = \Phi(\mathcal{X})^\circ \subset E'$. Since $\mathrm{span}\,[\Phi(\mathcal{X})] = \Psi(\mathcal{H}_0)$, it follows that $\mathrm{Ker}(\iota) = \Psi(\mathcal{H}_0)^\circ$. The conclusion now follows from the fact that for any subspace $F \subset E$, the image of F under $x \mapsto x'$ is dense in $(F^\circ)^\perp$. □

Proposition E.3. *Suppose $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ is a locally convex TVS and that $\mathcal{H}_0$ continuously embeds into $\mathcal{F}$. Then the map $\iota : E' \to \mathbb{R}^{\mathcal{X}}$ given by* (5) *maps E' continuously into $\mathcal{F}$. Furthermore*

- *Φ is universal to $\mathcal{F}$ iff the kernel k is universal to $\mathcal{F}$,*
- *Φ is characteristic to $\mathcal{F}'$ iff the kernel k is characteristic to $\mathcal{F}'$.*

Proof. S Substituting $\mathcal{F}$ by its completion if necessary, we may assume $\mathcal{F}$ is complete. Write $F_0 := \Psi(\mathcal{H}_0)' \subset E'$ and let $F \subset E'$ denote the closure of F_0 in E'. By Point (3) of Theorem E.2, it holds that

$$\mathrm{id}(\mathcal{H}_0) = \iota(F_0) \subset \mathcal{F}, \tag{20}$$

so by the assumption that $\mathrm{id} : \mathcal{H}_0 \hookrightarrow \mathcal{F}$ is continuous, the restriction $\iota|_{F_0} : F_0 \to \mathcal{F}$ is continuous. By definition of ι, it is easy to see that the restriction $\iota|_F$ agrees with the unique continuous extension of $\iota|_{F_0}$ to F. Hence $\iota\,|_F \colon F \to \mathcal{F}$ is continuous. We now write $E' = F \oplus F^\perp$. By Point (4) of Theorem E.2, we have $F = \mathrm{Ker}(\iota)^\perp$.

Note that $\mathrm{Ker}(\iota)$ is closed in the weak topology of E', and thus, *a fortiori*, under the strong (norm) topology. Indeed, if $\ell_n \to \ell$ pointwise in E', then for all $x \in \mathcal{X}$, $\lim_{n\to\infty} \ell_n(\Phi(x)) = \ell(\Phi(x))$. In particular, if $\ell_n \in \mathrm{Ker}(\iota)$ for all $n \geq 1$, then $\ell \in \mathrm{Ker}(\iota)$.

As a consequence, it holds that $\mathrm{Ker}(\iota) = F^\perp$, and thus

$$E' = F \oplus \mathrm{Ker}(\iota), \tag{21}$$

from which the continuity of $\iota : E' \to \mathcal{F}$ follows.

Finally, by (21) and the continuity of ι, the closures in $\mathcal{F}$ of $\iota(F_0)$ and $\iota(E')$ coincide. Combining with (20), it holds that $\iota(E')$ is dense in $\mathcal{F}$ iff $\mathrm{id}(\mathcal{H}_0)$ is dense in $\mathcal{F}$, from which the first equivalence follows. The second equivalence now follows from Theorem 2.3 and Proposition 7.2. □

Appendix F. Proof of Proposition 7.6

We continue the proof of Proposition 7.6 started in Section 7.2.

Claim 1. *It holds that Φ maps $\mathcal{X}$ continuously into the unit sphere of E.*

Proof. For every $x \in \mathcal{X}$, we have $\|\Phi(x)\|_E^2 = \sum_{q\in Q} |f_q(x)|^2 = \sum_{q\in Q} w(q)^2 = 1$. To show that Φ is continuous, let $y_m \to y \in \mathcal{X}$ as $y \to \infty$. Then

$$\|\Phi(y) - \Phi(y_m)\|_E^2 = \sum_{q\in Q} w(q)^2 \Big| \exp\Big(i\langle \sum_n q_n x_n, y\rangle\Big) - \exp\Big(i\langle \sum_n q_n x_n, y_m\rangle\Big)\Big|^2 ,$$

which converges to zero as $m \to \infty$ since w is square summable and $\lim_{m\to\infty}\langle \sum_n q_n x_n, y_m\rangle = \langle \sum_n q_n x_n, y\rangle$ for every fixed $q \in Q$. □

Claim 2. *The map $\iota : E' = E \to \mathbb{C}^{\mathcal{X}}$ given by* (5) *maps E continuously and densely into $\mathcal{F}$. In particular, Φ is universal to $\mathcal{F}$.*

Proof. Let E_0 denote the dense subspace of E consisting of function which eventually vanish. The image of $f \in E_0$ under ι is precisely

$$\langle \Phi(\cdot), f\rangle_E = \sum_{q\in Q} f_q(\cdot)\overline{f(q)},$$

hence ι manifestly maps E_0 surjectively onto $\mathcal{F}_0$. Moreover, since Φ maps $\mathcal{X}$ into the unit sphere of E, it follows by the Cauchy–Schwarz inequality that

$$|\iota(f)|_\infty = \sup_{x\in\mathcal{X}} |\langle \Phi(x), f\rangle_E| \leq |f|_E.$$

Hence $\iota : E_0 \to \mathcal{F}_0$ is continuous (even of unit norm). It readily follows that $\iota : E \to \mathcal{F}$ is continuous and dense. □

The kernel k associated to Φ is given by

$$k : \mathcal{X}\times\mathcal{X} \to \mathbb{C}, \quad k(x,y) = \langle \Phi(x), \Phi(y)\rangle_E = \sum_{q\in Q} f_q(x)\overline{f_q(y)} = \sum_{q\in Q} w(q)^2 \exp\Big(i\langle \sum_n q_n x_n, x - y\rangle_{\mathcal{X}}\Big).$$

Following the construction in Section 7.1, $\mathcal{H}_0 := \mathrm{span}\,(k_x \mid x \in \mathcal{X})$ can be identified with a subspace of E. The induced inner product takes the form $\langle k_x, k_y\rangle_{\mathcal{H}_0} = k(x,y)$.

By Claim 2, Point (3) of Theorem E.2, and first equivalence in Proposition E.3, we see that $\mathcal{H}_0$ is a subspace of $\mathcal{F}$ and the embedding $\iota : \mathcal{H}_0 \hookrightarrow \mathcal{F}$ maps $\mathcal{H}_0$ continuously and densely into $\mathcal{F}$. In particular k is characteristic to $\mathcal{F}'$, which concludes the proof of Point (1).

Claim 3. *The space of probability measures on $\mathcal{X}$ embeds injectively into $\mathcal{F}'$.*

Proof. Clearly every probability measure on $\mathcal{X}$ is an element of $\mathcal{F}'$. It remains to show that if $\mu(f) = \nu(f)$ for all $f \in \mathcal{F}$, then $\mu = \nu$.

Observe that $\bar{f}_q = w(q)f_{-q}/w(-q)$ and $f_q f_r = w(q)w(r)f_{q+r}/w(q+r)$. It follows that $\mathcal{F}_0$ is closed under multiplication and complex conjugation. We now claim that $\mathcal{F}_0$ separates points. Indeed, for some $x, y \in \mathcal{X}$, suppose that $f_q(x) = f_q(y)$ for all $q \in Q$. This is equivalent to

$$\exp\big(i\langle \sum_k q_k x_k, x\rangle_{\mathcal{X}}\big) = \exp\big(i\langle \sum_k q_k x_k, y\rangle_{\mathcal{X}}\big), \quad \forall q \in Q,$$

from which it follows that $x = y$. The conclusion now follows by Stone-Weierstrass and a compactification argument (see, e.g., Theorem 2.6 or [5, Ex. 7.14.79]). □

Point (2) now follows from Claim 3. In particular, k is characteristic to probability measures on $\mathcal{X}$. Finally, we have the following claim which proves Point (3) and concludes the proof of Proposition 7.6.

For a probability measures μ on $\mathcal{X}$, let $\Phi_k(\mu)$ be the unique element in $\mathcal{H}$ for which $\mu(h) = \langle h, \Phi_k(\mu)\rangle_{\mathcal{H}}$ (note that $\Phi_k(\mu)$ clearly exists since μ is a continuous linear functional on $\mathcal{H}$ via the map $\iota : \mathcal{H} \to \mathcal{F}$).

Claim 4. *The distance $d_k(\mu, \nu) := |\Phi_k(\mu) - \Phi_k(\nu)|_{\mathcal{H}}$ does not metrize the weak topology on probability measures on $\mathcal{X}$.*

Proof. For $x \in \mathcal{X}$, consider the Dirac delta probability measures δ_x. Then for $h = \sum a_n k_{y_n} \in \mathcal{H}_0$, it holds that

$$\delta_x(h) = h(x) = \sum a_n k(y_n, x) = \langle h, k_x\rangle,$$

from which it follows that $\Phi_k(\delta_x) = k_x \in \mathcal{H}_0$. For the orthonormal basis $x_1, x_2, \ldots$, consider the sequence of probability measures $\delta_{x_1}, \delta_{x_2}, \ldots$. It holds that

$$\begin{aligned} |k_{x_n} - k_0|^2_{\mathcal{H}} &= 2 - 2\mathfrak{R}\left[k(x_n, 0)\right] \\ &= 2 - 2\mathfrak{R}\Big[\sum_{q\in Q} e^{iq_n} w(q)^2\Big] \\ &\leq 2\sum_{q\in Q_n} w(q)^2, \end{aligned}$$

where $Q_n \subset Q$ is the subset of all $q = (q_1, q_2, \ldots)$ for which $q_n \neq 0$. However, by construction of Q and the square-summability of w, we see that $\sum_{q\in Q_n} w(q)^2 \to 0$ as $n \to \infty$. It follows that

$$\lim_{n\to\infty} d_k(\delta_{x_n}, \delta_0) = 0.$$

However, it clearly holds that the sequence of probability measures $(\delta_{x_n})_{n\geq 1}$ does not converge weakly to δ_0 (and is not even tight). □

Appendix G. Computation: truncation and discretization

Following the discussion in Section 7.5, in practice we usually work with the truncated space $\prod_{m=0}^{M} V^{\otimes m}$ rather than the full space $\mathbf{T}(V)$. Furthermore, we usually only have access to a finite collection of time points $(x(t))_{t\in\pi}$ for some partition π. The purpose of this appendix is to collect explicit convergence results for the corresponding normalized signature feature map and kernel. We first provide an approximation scheme for geometric p-rough paths and then an alternative scheme for the special case of bounded variation paths.

In what follows, let $\psi : [1, \infty) \to [1, \infty)$ be a function with $\psi(1) = 1$ satisfying the conditions of item (iii) of Proposition A.2 and $\Lambda : \mathbf{T}_1(V) \to \mathbf{T}_1(V)$ the corresponding map.

Definition G.1. *Let V be a Banach space, $p \geq 1$, $\mathbf{x} \in C^p$, and $N \geq \lfloor p \rfloor$. Define $\mathrm{S}_N(\mathbf{x})$ as the canonical projection of $\mathrm{S}(\mathbf{x}) \in \mathbf{T}(V)$ to $\prod_{m=0}^{N} V^{\otimes m}$.*

Further let $M \geq N$ and $\pi = \{0 = t_0 < \ldots < t_n = 1\}$ a partition of $[0, 1]$. Define[13] $\mathrm{S}_{N,i}(\mathbf{x}) := \mathrm{S}_N\left(\mathbf{x}|_{[t_{i-1},t_i]}\right)$ *and*

$$\mathrm{S}^{\pi}_{M,N}(\mathbf{x}) := \mathrm{S}_{N,1}(\mathbf{x}) \otimes \ldots \otimes \mathrm{S}_{N,n}(\mathbf{x}) \in \prod_{m=0}^{M} V^{\otimes m},$$

where we identify $\mathrm{S}_{N,i}(\mathbf{x})$ canonically as an element of $\prod_{m=0}^{M} V^{\otimes m}$ and recall the product $\otimes$ on $\prod_{m=0}^{M} V^{\otimes m}$ from Appendix C. Further, define

$$\Phi^{\pi}_{M,N} : C^p \to \prod_{m=0}^{M} V^{\otimes m}, \quad \Phi^{\pi}_{M,N}(\mathbf{x}) := \Lambda \circ \mathrm{S}^{\pi}_{M,N}(\mathbf{x}).$$

[13] We introduced S in Theorem C.5 for paths parameterized by $[0, 1]$, but since Theorem C.5 implies parameterization invariance of S, we can canonically replace $[0, 1]$ with an arbitrary interval $[t_{i-1}, t_i]$.

where we canonically identified $\prod_{m=0}^{M} V^{\otimes m}$ *as a subspace of* $\mathbf{T}(V)$ *via* $(\mathbf{t}^0, \dots, \mathbf{t}^M) \mapsto (\mathbf{t}^0, \dots, \mathbf{t}^M, 0, \dots)$ *which is closed under the dilation map* δ_λ *so that* $\Lambda \circ \mathrm{S}^\pi_{M,N}(\mathbf{x})$ *is well-defined as an element of* $\prod_{m=0}^{M} V^{\otimes m}$.

Remark G.2. $\mathrm{S}^\pi_{M,N}(\mathbf{x})$ *is the level-N Euler scheme of a linear rough differential equation (RDE) in* $\prod_{m=0}^{M} V^{\otimes m}$ *driven by* $\mathbf{x}$*. In particular, setting* $N = 1$ *yields the approximations discussed in* [33, Sec. 4.1].

Proposition G.3. *Let* $V = \mathbb{R}^d$ *equipped with an inner product,* $p \geq 1$*, and* $M \geq N \geq \lfloor p \rfloor$.

(1) There exists a constant $C > 0$ *such that for any* $\mathbf{x} \in C^p(\mathbb{R}^d)$ *and any partition* π *of* $[0,1]$ *for which* $\|\mathbf{x}^\pi\|_{0,p\text{-var}} := \max_{t_i \in \pi} \|\mathbf{x}\|_{p\text{-var};[t_{i-1},t_i]} \leq 1$*, it holds that*

$$\|\mathrm{S}^\pi_{M,N}(\mathbf{x}) - \mathrm{S}_M(\mathbf{x})\| \leq C(\|\mathbf{x}\|^{p+2}_{p\text{-var}} \vee \|\mathbf{x}\|^{p+2M}_{p\text{-var}})\|\mathbf{x}^\pi\|^{N+1-p}_{0,p\text{-var}} .$$

(2) Let $\overline{K} := \|\psi\|^{1/2}_\infty(1 + K^{1/2} + 2\|\psi\|^{1/2}_\infty)$*. Let* $\mathbf{y} \in C^p(\mathbb{R}^d)$ *and* π' *another partition of* $[0,1]$*. Recall the kernel* $\mathrm{k}_M(\mathbf{x}, \mathbf{y})$ *from Section 7.5. Then*

$$\begin{aligned}\left|\mathrm{k}_M(\mathbf{x},\mathbf{y}) - \langle \Phi^\pi_{M,N}(\mathbf{x}), \Phi^{\pi'}_{M,N}(\mathbf{y})\rangle\right| \leq \overline{K}(\|\mathrm{S}^\pi_{M,N}(\mathbf{x}) - \mathrm{S}_M(\mathbf{x})\|^{1/2} \vee \|\mathrm{S}^\pi_{M,N}(\mathbf{x}) - \mathrm{S}_M(\mathbf{x})\| \\ + \|\mathrm{S}^\pi_{M,N}(\mathbf{y}) - \mathrm{S}_M(\mathbf{y})\|^{1/2} \vee \|\mathrm{S}^\pi_{M,N}(\mathbf{y}) - \mathrm{S}_M(\mathbf{y})\|) .\end{aligned}$$

Proof. Note that (2) follows from items (i) and (iii) of Proposition A.2 and the Cauchy–Schwarz inequality. It remains to show (1). Since the case $M = \lfloor p \rfloor$ is trivial, we may suppose $M > \lfloor p \rfloor$. Denoting $\pi = \{0 = t_0 < \dots < t_n = 1\}$, we note that by Chen's identity, $\mathrm{S}_M = \mathrm{S}_{M,1} \otimes \dots \otimes \mathrm{S}_{M,n}$ (where we drop the reference to $\mathbf{x}$). Applying a telescoping sum and sub-multiplicativity of the norm $\|\cdot\|$ on $\prod_{m=0}^{M} (\mathbb{R}^d)^{\otimes m}$, we obtain

$$\|\mathrm{S}_{M,N} - \mathrm{S}_M\| \leq \sum_{i=1}^{n} \|\mathrm{S}_{N,1} \dots \mathrm{S}_{N,i-1}\| \, \|\mathrm{S}_{N,i} - \mathrm{S}_{M,i}\| \, \|\mathrm{S}_{M,i+1} \dots \mathrm{S}_{M,n}\| .$$

Denoting $\omega(s,t) := \|\mathbf{x}\|^p_{p\text{-var};[s,t]}$ and $\lesssim$ a bound with proportionality constant independent of $\mathbf{x}$ and π, we have for all $i = 1, \dots, n$

$$\begin{aligned}\|\mathrm{S}_{N,i} - \mathrm{S}_{M,i}\| &\lesssim \omega(t_{i-1}, t_i)^{(N+1)/p} \vee \omega(t_{i-1}, t_i)^{M/p} , \\ \|\mathrm{S}_{M,i+1} \dots \mathrm{S}_{M,n}\| &\lesssim \omega(0,1)^{1/p} \vee \omega(0,1)^{M/p} , \\ \|\mathrm{S}_{N,1} \dots \mathrm{S}_{N,i-1}\| &\lesssim \omega(0,1)^{1/p} \vee \omega(0,1)^{M/p}\end{aligned}$$

(the first two bounds are obvious, while the last follows from [14, Lem. 4.16] applied to the homogeneous group $\{\mathbf{t} \in \prod_{m=0}^{M} \mathbb{R}^{d^{\otimes m}} : \mathbf{t}^0 = 1\}$). Using the assumption $\|\mathbf{x}^\pi\|_{0,p\text{-var}} \leq 1$ and super-additivity of $\omega^{(N+1)/p}$, it follows that

$$\begin{aligned}\|\mathrm{S}_{M,N}(\mathbf{x}) - \mathrm{S}_M(\mathbf{x})\| &\lesssim \left(\omega(0,1)^{2/p} \vee \omega(0,1)^{2M/p}\right) \sum_{i=1}^{n} \omega(t_{i-1}, t_i)^{(N+1)/p} \\ &\leq (\|\mathbf{x}\|^{p+2}_{p\text{-var}} \vee \|\mathbf{x}\|^{p+2M}_{p\text{-var}})\|\mathbf{x}^\pi\|^{N+1-p}_{0,p\text{-var}} ,\end{aligned}$$

which proves (1). □

We state now a variant of Proposition G.3 in the case $p = 1$, which is essentially contained in [33]. The point here is the explicit error bounds (independent of V, M, N) obtained by exploiting the bounded variation nature of the paths.

Definition G.4. *Let* V *be a Banach space,* $M \geq N \geq 1$*,* $x \in C([0,1], V)$*, and* $\pi = \{0 = t_0 < \dots < t_n = 1\}$ *a partition of* $[0,1]$*. Define*

$$\mathrm{S}^{\mathrm{linear},\pi}_{M,N}(x) := \left(\frac{(x(t_1) - x(t_0))^{\otimes m}}{m!}\right)_{m=0}^{N} \otimes \dots \otimes \left(\frac{(x(t_n) - x(t_{n-1}))^{\otimes m}}{m!}\right)_{m=0}^{N} \in \prod_{m=0}^{M} V^{\otimes m},$$

where we identify $((x(t_i) - x(t_{i-1}))^{\otimes m}/m!)_{m=0}^{N}$ *canonically with an element of* $\prod_{m=0}^{M} V^{\otimes m}$*. Further, define*

$$\Phi^{\mathrm{linear},\pi}_{M,N}(x) := \Lambda \circ \mathrm{S}^{\mathrm{linear},\pi}_{M,N}(x)$$

Remark G.5. *The tensor* $\mathrm{S}^{\mathrm{linear},\pi}_{M,N}(x)$ *is the level-N Euler approximation of the level-M signature of the piecewise linear interpolation of* x *along the partition* π*; this is precisely the approximation discussed in* [33, Appx. B]*. In particular,* $\mathrm{S}^{\mathrm{linear},\pi}_{M,1}(x) = \mathrm{S}^\pi_{M,1}(x)$*, while* $\mathrm{S}^{\mathrm{linear},\pi}_{M,M}(x)$ *is the full level-M signature of the piecewise linear interpolation.*

If x admits a lift $\mathbf{x} \in C^p$, then, in contrast to $\Phi^\pi_{M,N}$, we do not in general expect an (asymptotic) improvement in the approximation $\Phi_M(\mathbf{x}) \approx \Phi^{\mathrm{linear},\pi}_{M,N}$ from larger values of N. This is due to the fact that $\mathrm{S}^{\mathrm{linear},\pi}_{M,N}$ ignores the higher order iterated integrals of $\mathbf{x}$ (in fact, without additional information on $\mathbf{x}$, we cannot expect $\Phi^{\mathrm{linear},\pi}_{M,N}$ to converge as the mesh of π tends to zero). However, we still obtain the following convergence result for bounded variation paths.

Proposition G.6. *Let V be a Hilbert space, $x \in C^1(V)$, and π a partition of $[0,1]$.*

(1) For all $N = 1, \dots, M$ it holds that

$$\| \mathrm{S}^{\mathrm{linear},\pi}_{M,N}(x) - \mathrm{S}_M(x)\| \leq \|x\|_{1-\mathrm{var}} e^{\|x\|_{1-\mathrm{var}}} \|x^\pi\|_{0,1-\mathrm{var}} ,$$

where $\|x^\pi\|_{0,1-\mathrm{var}} := \max_{t_i \in \pi} \|x\|_{1-\mathrm{var};[t_{i-1},t_i]}$.

(2) Let $\overline{K} := \|\psi\|_\infty^{1/2}(1 + K^{1/2} + 2\|\psi\|_\infty^{1/2})$. If $y \in C^1(V)$ and π' is another partition of $[0,1]$, then for all $N = 1, \dots, M$

$$\begin{aligned}\left|\mathrm{k}_M(x,y) - \langle \Phi^{\mathrm{linear},\pi}_{M,N}(x), \Phi^{\mathrm{linear},\pi}_{M,N}(y)\rangle\right| \leq \overline{K}(&\| \mathrm{S}^{\mathrm{linear},\pi}_{M,N}(x) - \mathrm{S}_M(x)\|^{1/2} \vee \| \mathrm{S}^{\mathrm{linear},\pi}_{M,N}(x) - \mathrm{S}_M(x)\| \\ &+ \| \mathrm{S}^{\mathrm{linear},\pi}_{M,N}(y) - \mathrm{S}_M(y)\|^{1/2} \vee \| \mathrm{S}^{\mathrm{linear},\pi}_{M,N}(y) - \mathrm{S}_M(y)\|) .\end{aligned}$$

Proof. Point (1) for $N = 1$ is a restatement of [33, Cor. 4.3], while the case $2 \leq N \leq M$ follows from the same proof given in [33, Appx. A]. Point (2) follows from items (i) and (iii) of Proposition A.2 as in the proof of Proposition G.3. □

Remark G.7. *Note that choosing large $N \geq 2$ does not confer benefit for the approximation $\mathrm{S}^{\mathrm{linear},\pi}_{M,N}(x) \approx \mathrm{S}_M(x)$. However large values of N lead to "more geometric" approximations (with $N = M$ being genuinely geometric since $\mathrm{S}^{\mathrm{linear},\pi}_{M,M}$ is the signature of a bounded variation path).*

Remark G.8. *Given as input the quantities*

$$\{\langle x(t_{i-1}) - x(t_i), y(t'_{i'-1}) - y(t'_{i'})\rangle : 1 \leq i \leq n, 1 \leq i' \leq n'\} ,$$

one can compute $\langle \mathrm{S}^{\mathrm{linear},\pi}_{M,N}(x), \mathrm{S}^{\mathrm{linear},\pi}_{M,N}(y)\rangle_{\prod_{m=0}^M V^{\otimes m}}$ using $O(N^2Mnn')$ operations and using $O(N^2nn')$ memory, see [33, Alg. 6].

References

[1] R. F. Bass, B. M. Hambly, and T. J. Lyons. Extending the Wong-Zakai theorem to reversible Markov processes. J. Eur. Math. Soc. (JEMS), 4(3):237–269, 2002.

[2] Alain Berlinet and Christine Thomas-Agnan. Reproducing kernel Hilbert spaces in probability and statistics. Springer Science & Business Media, 2011.

[3] H. Boedihardjo and I. Chevyrev. An isomorphism between branched and geometric rough paths. ArXiv e-prints, December 2017. To appear in Ann. Inst. H. Poincaré Probab. Statist.

[4] Horatio Boedihardjo, Xi Geng, Terry Lyons, and Danyu Yang. The signature of a rough path: uniqueness. Adv. Math., 293:720–737, 2016.

[5] V. I. Bogachev. Measure theory. Vol. I, II. Springer-Verlag, Berlin, 2007.

[6] Richard P. Brent. Algorithms for minimization without derivatives. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1973. Prentice-Hall Series in Automatic Computation.

[7] Y. Bruned, I. Chevyrev, P. K. Friz, and R. Preiss. A Rough Path Perspective on Renormalization. ArXiv e-prints, January 2017.

[8] R Creighton Buck. Bounded continuous functions on a locally compact space. The Michigan Mathematical Journal, 5(2):95–104, 1958.

[9] T. Cass and M. P. Weidner. Tree algebras over topological vector spaces in rough path theory. ArXiv e-prints, April 2016.

[10] Thomas Cass, Bruce K. Driver, Nengli Lim, and Christian Litterer. On the integration of weakly geometric rough paths. J. Math. Soc. Japan, 68(4):1505–1524, 2016.

[11] I. Chevyrev and P. K. Friz. Canonical RDEs and general semimartingales as rough paths. ArXiv e-prints, April 2017. To appear in Annals of Probabilty.

[12] I. Chevyrev, V. Nanda, and H. Oberhauser. Persistence paths and signature features in topological data analysis. ArXiv e-prints, June 2018.

[13] I. Chevyrev and M. Ogrodnik. A support and density theorem for Markovian rough paths. Electron. J. Probab., 23(56):16 pp., 2018.

[14] Ilya Chevyrev. Random walks and Lévy processes as rough paths. Probab. Theory Related Fields, 170(3-4):891–932, 2018.

[15] Ilya Chevyrev and Terry Lyons. Characteristic functions of measures on geometric rough paths. Ann. Probab., 44(6):4049–4082, 2016.

[16] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. JMLR: Workshop and Conference Proceedings, 2016.

[17] Laure Coutin and Zhongmin Qian. Stochastic analysis, rough path analysis and fractional Brownian motions. Probab. Theory Related Fields, 122(1):108–140, 2002.

[18] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. Bulletin of the American mathematical society, 39(1):1–49, 2002.

[19] Thomas Fawcett. Problems in stochastic analysis: connections between rough paths and non-commutative harmonic analysis. PhD thesis, University of Oxford, 2003.

[20] Peter Friz and Atul Shekhar. General rough integration, Lévy rough paths and a Lévy–Kintchine type formula. Ann. Probab., 45(4):2707–2765, 2017.

[21] Peter Friz and Nicolas Victoir. On uniformly subelliptic operators and stochastic area. Probab. Theory Related Fields, 142(3-4):475–523, 2008.

[22] Peter K. Friz, Benjamin Gess, Archil Gulisashvili, and Sebastian Riedel. The Jain-Monrad criterion for rough paths and applications to random Fourier series and non-Markovian Hörmander theory. Ann. Probab., 44(1):684–738, 2016.

[23] Peter K. Friz and Nicolas B. Victoir. Multidimensional stochastic processes as rough paths: theory and applications. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2010.

[24] Peter K. Friz and Huilin Zhang. Differential equations driven by rough paths with jumps. J. Differential Equations, 264(10):6226–6301, 2018.

[25] Robin Giles. A generalization of the strict topology. Transactions of the American Mathematical Society, 161:467–474, 1971.

[26] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. Journal of Machine Learning Research, 13(Mar):723–773, 2012.
[27] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In Advances in neural information processing systems, pages 673–681, 2009.
[28] Massimiliano Gubinelli. Ramification of rough paths. Journal of Differential Equations, 248(4):693–721, 2010.
[29] Martin Hairer and David Kelly. Geometric versus non-geometric rough paths. Ann. Inst. Henri Poincaré Probab. Stat., 51(1):207–251, 2015.
[30] Ben Hambly and Terry Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. Ann. of Math. (2), 171(1):109–167, 2010.
[31] Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In Advances in Neural Information Processing Systems, pages 181–189, 2016.
[32] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.
[33] F. J Király and H. Oberhauser. Kernels for sequentially ordered data. ArXiv e-prints, January 2016.
[34] Antoine Lejay. Stochastic differential equations driven by processes generated by divergence form operators. I. A Wong-Zakai theorem. ESAIM Probab. Stat., 10:356–379 (electronic), 2006.
[35] Terry Lyons. Differential equations driven by rough signals. Rev. Mat. Iberoamericana, 14(2):215–310, 1998.
[36] Terry Lyons and Zhongmin Qian. System control and rough paths. Oxford Mathematical Monographs. Oxford University Press, Oxford, 2002. Oxford Science Publications.
[37] Terry J. Lyons, Michael Caruana, and Thierry Lévy. Differential equations driven by rough paths, volume 1908 of Lecture Notes in Mathematics. Springer, Berlin, 2007. Lectures from the 34th Summer School on Probability Theory held in Saint-Flour, July 6–24, 2004, With an introduction concerning the Summer School by Jean Picard.
[38] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. Foundations and Trends® in Machine Learning, 10(1-2):1–141, 2017.
[39] Alfred Müller. Integral probability metrics and their generating classes of functions. Advances in Applied Probability, 29(2):429–443, 1997.
[40] Hao Ni. The expected signature of a stochastic process. PhD thesis, University of Oxford, 2012.
[41] Anastasia Papavasiliou, Christophe Ladroue, et al. Parameter estimation for rough differential equations. The Annals of Statistics, 39(4):2047–2073, 2011.
[42] Svetlozar Todorov Rachev. Probability metrics and the stability of stochastic models, volume 269. John Wiley & Son Ltd, 1991.
[43] Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry A Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In AAAI, pages 3571–3577, 2015.
[44] Bernhard Schölkopf and Alexander J Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.
[45] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, et al. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. The Annals of Statistics, 41(5):2263–2291, 2013.
[46] C.-J. Simon-Gabriel and B. Schölkopf. Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions. ArXiv e-prints, April 2016.
[47] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. Journal of Machine Learning Research, 11(Apr):1517–1561, 2010.
[48] Stephen Willard. General topology. Courier Corporation, 1970.
[49] David R. E. Williams. Path-wise solutions of stochastic differential equations driven by Lévy processes. Rev. Mat. Iberoamericana, 17(2):295–329, 2001.